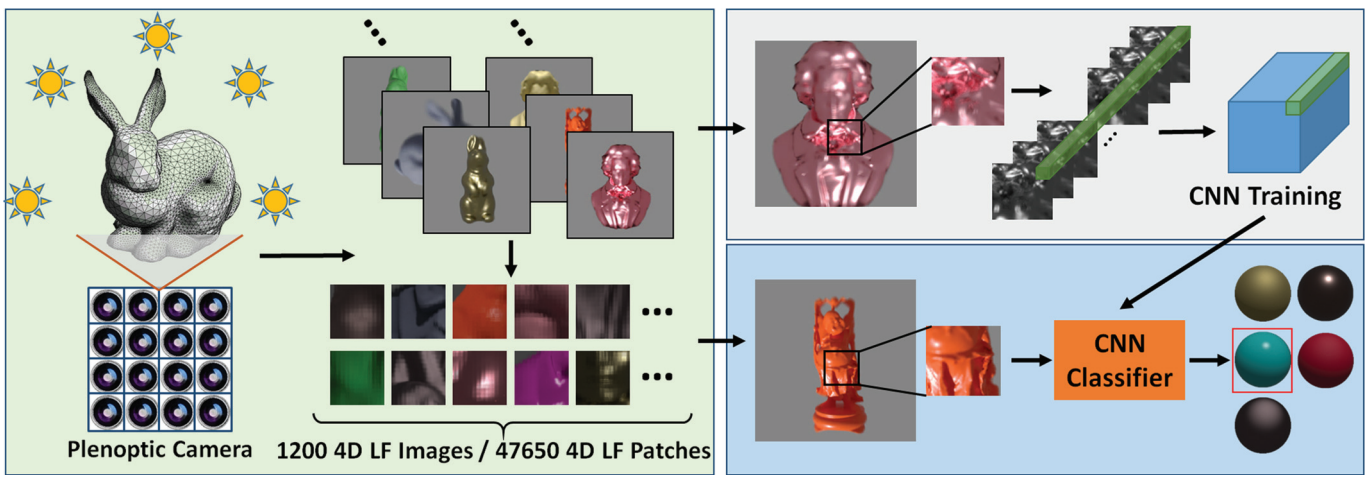
IEEE JOURNAL OF
SELECTED TOPICS IN SIGNAL PROCESSING

IEEE Signal Processing Society

www.ieee.org/sp/index.html

OCTOBER 2017	VOLUME 11	NUMBER 7	IJSTGY	(ISSN 1932-4553)
DECEMBER 2017	VOLUME 11	NUMBER 8		

ISSUE ON LIGHT FIELD IMAGE PROCESSING



An overview of the proposed dataset and method for identification of BRDF. Light field images allow simultaneous spatial and angular sampling of an object, which provide rich information for BRDF identification. For more, see "Identifying Surface BRDF From a Single 4-D Light Field Image via Deep Neural Network" by F. Lu *et al.,* p. 1047.

For the October 2017 issue, see p. 921 for Table of Contents.
For the December 2017 issue, see p. 1235 for Table of Contents.

IEEE

IEEE SIGNAL PROCESSING SOCIETY

The Signal Processing Society is an organization, within the framework of the IEEE, of members with principal professional interest in the technology of transmission, recording, reproduction, processing, and measurement of speech; other audio-frequency waves and other signals by digital electronic, electrical, acoustic, mechanical, and optical means; the components and systems to accomplish these and related aims; and the environmental, psychological, and physiological factors of these technologies. All members of the IEEE are eligible for membership in the Society and will receive this TRANSACTIONS upon payment of the annual Society membership fee of $20.00 plus an annual subscription fee of $185.00. For information on joining, write to the IEEE at the address below. *Member copies of Transactions/Journals are for personal use only*.

J-STSP Senior Editorial Board

SHRIKANTH (SHRI) S. NARAYANAN, *Editor-in-Chief*
University of Southern California
Los Angeles, CA 90089 USA

B. ATAL
University of Washington
Seattle, WA 98195 USA

C. BOUMAN
Purdue University
W. Lafayette, IN 47907 USA

A. BOVIK
University of Texas
Austin, TX 78712 USA

R. CHELLAPPA
University of Maryland
College Park, MD 20742 USA

M.-G. DI BENEDETTO
University of Rome, La Sapienza
00185 Rome, Italy

P. DINIZ
Universidade Federal do Rio de Janeiro
21941-972 - Rio de Janeiro - RJ, Brazil

Y. ELDAR
Technion–Israel Institute of Technology
Haifa, 3200003, Israel

G. GIANNAKIS
University of Minnesota
Minneapolis, MN 55455 USA

M. S. GRECO
University of Pisa
56122 Pisa, Italy

A. Hero
University of Michigan–Ann Arbor
Ann Arbor, MI 48109 USA

C.-C. JAY KUO
University of Southern California
Los Angeles, CA 90089-2564 USA

H. MENG
Chinese University of Hong Kong
Hong Kong, China

M. OSTENDORF
University of Washington
Seattle, WA 98195-2500 USA

V. POOR
Princeton University
Princeton, NJ 08544 USA

B. RAO
University of California at San Diego
La Jolla, CA 92093 USA

IEEE Officers

KAREN BARTLESON, *President*
JAMES A. JEFFERIES, *President-Elect*
WILLIAM P. WALSH, *Secretary*
JOHN W. WALZ, *Treasurer*
BARRY L. SHOOP, *Past President*
S. K. RAMESH, *Vice President, Educational Activities*
SAMIR M. EL-GHAZALY, *Vice President, Publication Services and Products*
MARY ELLEN RANDALL, *Vice President, Member and Geographic Activities*
FORREST D. WRIGHT, *President, Standards Association*
MARINA RUGGIERI, *Vice President, Technical Activities*
KAREN S. PEDERSEN, *President, IEEE-USA*

CELIA L. DESMOND, *Director, Division III—Communications Technology*

IEEE Executive Staff

DR. E. JAMES PRENDERGAST, *Executive Director & Chief Operating Officer*
THOMAS SIEGERT, *Business Administration*
ELENA GERSTMANN, *Corporate Activities*
DOUGLAS GORHAM, *Educational Activities*
EILEEN M. LACH, *General Counsel & Corporate Compliance Officer*
VACANT, *Human Resources*
CHRIS BRANTLEY, *IEEE-USA*
CHERIF AMIRAT, *Information Technology*
KAREN HAWKINS, *Marketing*
CECELIA JANKOWSKI, *Member and Geographic Activities*
MICHAEL FORSTER, *Publications*
KONSTANTINOS KARACHALIOS, *Standards Association*
MARY WARD-CALLAN, *Technical Activities*

IEEE Periodicals
Transactions/Journals Department

Senior Director, Publishing Operations: FRAN ZAPPULLA
Director, Editorial Services: DAWN MELLEY *Director, Production Services:* PETER M. TUOHY
Associate Director, Editorial Services: JEFFREY E. CICHOCKI *Associate Director, Information Conversion and Editorial Support:* KEVIN LISANKIE
Managing Editor: MARTIN J. MORAHAN *Journals Coordinator:* ANDREW SWARTZ

Digital Object Identifier 10.1109/JSTSP.2017.2736331

IEEE JOURNAL OF
SELECTED TOPICS IN SIGNAL PROCESSING

IEEE Signal Processing Society

www.ieee.org/sp/index.html

OCTOBER 2017 VOLUME 11 NUMBER 7 IJSTGY (ISSN 1932-4553)

ISSUE ON LIGHT FIELD IMAGE PROCESSING

EDITORIAL

Introduction to the Issue on Light Field Image Processing ..... *Y. Liu, L. Fang, D. Gutierrez, Q. Wang, J. Yu, and F. Wu* 923

PAPERS

Light Field Image Processing: An Overview ..... *G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu* 926

Depth Estimation From Light Field by Accumulating Binary Maps Based on Foreground–Background Separation ..... *J. Y. Lee and R.-H. Park* 955

Occlusion-Model Guided Antiocclusion Depth Estimation in Light Field ..... *H. Zhu, Q. Wang, and J. Yu* 965

Seeing Beyond Foreground Occlusion: A Joint Framework for SAP-Based Scene Depth and Appearance Reconstruction ..... *Z. Xiao, L. Si, and G. Zhou* 979

Single Aperture Spectral+ToF Compressive Camera: Toward Hyperspectral+Depth Imagery ..... *H. Rueda, C. Fu, D. L. Lau, and G. R. Arce* 992

From the Calibration of a Light-Field Camera to Direct Plenoptic Odometry ..... *N. Zeller, F. Quint, and U. Stilla* 1004

(Contents Continued on Page 922)

IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING (ISSN 1932-4553) is published eight times a year in February, March, April, June, August, September, October, and December by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $420.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee of $31.00 is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141. Copyright © 2017 by the Institute of Electrical and Electronics Engineers, Inc. All rights reserved. **Postmaster:** Send address changes to IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: PitneyBowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ONM5W3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE

(Contents Continued from Page 921)

Characterization of Plenoptic Imaging Systems and Efficient Volumetric Estimation From Plenoptic Data ..... *P. Anglin, S. J. Reeves, and B. S. Thurow* 1020
A Two-Phase 3-D Reconstruction Approach for Light Microscopy Axial-View Imaging ..... *Y. Guo, Y. Zhang, and F. J. Verbeek* 1034
Identifying Surface BRDF From a Single 4-D Light Field Image via Deep Neural Network ..... *F. Lu, L. He, S. You, X. Chen, and Z. Hao* 1047
Super Resolution of Light Field Images Using Linear Subspace Projection of Patch-Volumes ..... *R. A. Farrugia, C. Galea, and C. Guillemot* 1058
Image Reconstruction via Manifold Constrained Convolutional Sparse Coding for Image Sets ..... *L. Yang, C. Li, J. Han, C. Chen, Q. Ye, B. Zhang, X. Cao, and W. Liu* 1072
Accelerated Shearlet-Domain Light Field Reconstruction ..... *S. Vagharshakyan, R. Bregovic, and A. Gotchev* 1082
Comparison and Evaluation of Light Field Image Coding Approaches ..... *I. Viola, M. Řeřábek, and T. Ebrahimi* 1092
Pseudo-Sequence-Based 2-D Hierarchical Coding Structure for Light-Field Image Compression ..... *L. Li, Z. Li, B. Li, D. Liu, and H. Li* 1107
Light Field Image Coding Using High-Order Intrablock Prediction ..... *R. J. S. Monteiro, P. J. L. Nunes, N. M. M. Rodrigues, and S. M. M. Faria* 1120
Light Field Compression With Homography-Based Low-Rank Approximation ..... *X. Jiang, M. L. Pendu, R. A. Farrugia, and C. Guillemot* 1132
Minimum Description Length Sparse Modeling and Region Merging for Lossless Plenoptic Image Compression ..... *P. Helin, P. Astola, B. Rao, and I. Tabus* 1146
Analyzing Interfaces and Workflows for Light Field Editing ..... *M. Ortín, A. Jarabo, B. Masia, and D. Gutierrez* 1162
Image Reshaping for Efficient Compression of Plenoptic Content ..... *X. Jin, H. Han, and Q. Dai* 1173
Superrays for Efficient Light Field Processing ..... *M. Hog, N. Sabater, and C. Guillemot* 1187
Designing Three-Dimensional Cellular Automata Based Video Authentication With an Optical Integral Imaging Generated Memory-Distributed Watermark ..... *X.-W. Li, S.-T. Kim, and Q.-H. Wang* 1200
Quantifying Spatial and Angular Resolution of Light-Field 3-D Displays ..... *P. T. Kovács, R. Bregović, A. Boev, A. Barsi, and A. Gotchev* 1213
Layered Display with Accommodation Cue Using Scattering Polarizers ..... *S. Moon, C.-K. Lee, D. Lee, C. Jang, and B. Lee* 1223

Introduction to the Issue on Light Field Image Processing

AS A representative approach for passive depth sensing, light field imaging and processing techniques have recently attracted a lot of attention from the computer vision, computer graphics, and signal processing communities. Compared with 2D images, light fields are 4D structures that encode also angular information, providing ways for better vision and image understanding, as well as opening new ways to interact with such information. Traditional vision problems like segmentation, scene parsing, salience detection, video stabilization, object detection, tracking and recognition, etc. may be better solved leveraging the additional information encoded in light fields. Moreover, they also pave the way for new applications such as novel displays, light field microscopy, etc.

Nevertheless, light field technology still suffers from crucial problems such as low spatial resolution, poor reconstructed depth quality, or large data storage requirements. Similarly, human interaction with four-dimensional data remains a challenging problem. Therefore, light field research is still in need of better theories and methods, targeting for more efficient data capture, analysis, processing, and storage. To this end, this special issue aims at providing readers with the latest developments and emerging technologies that aim to advance the field, and in the end make light fields more practical. For this purpose, 23 papers have been selected during the review process from 43 submissions.

The first paper, entitled "Light Field Image Processing: An Overview," by Wu *et al.* provides a timely and comprehensive overview of the field, including recent advances. The authors review and discuss key technologies, challenges, and the future development trends, including the basic representation and theory of light fields, acquisition, super-resolution, depth estimation, compression, editing, display, interaction, and of course, applications. This paper will help novel researchers to quickly understand the state of the art of light field research, while more experienced ones will have a one-stop reference document.

The most important advantage provided by light field data is its encoded angular information for depth estimation. The next four papers are devoted to depth acquisition from light field data. In the paper entitled "Depth Estimation from Light Fields by Accumulating Binary Maps Based on Foreground-Background Separation", Lee and Park model the depth estimation by separating the foreground and background with respect to a series of focused planes. The final depth map is obtained by accumulating the resulting binary maps. To handle occlusions, Zhu *et al.* present a complete occlusion model in their paper "Occlusion-Model Guided Anti-Occlusion Depth Estimation in Light Field". The model explores both single- and multi-occluders, deriving the occluder consistency between the spatial and angular dimensions. To recover the target scene depth in the presence of complex foreground occlusions, Xiao *et al.* combine synthetic aperture photography with depth estimation in their paper "Seeing beyond foreground occlusion: A joint framework for SAP-based scene depth and appearance reconstruction". The initial depth map is estimated by selecting optimal views with pixel-based clustering. Then a joint optimization model for depth recovery and appearance reconstruction is presented. Rather than estimating depth of a captured scene, Rueda *et al.* develop a depth imaging camera in "Single Aperture Spectral+ToF Compressive Camera: Towards Hyperspectral+Depth Imagery", combining a spectral imager and a Time-of-Flight (ToF) sensor. In their prototype, they use a digital micro-mirror device to emulate coded apertures, and a ToF sensor to integrate the incoming light field.

The angular information of light fields can be further exploited for 3D reconstruction of both macro- and micro-scale scenes. The next four papers deal with 3D scene reconstruction of different kinds of scenes and targets. In "From the Calibration of a Light-Field Camera to Direct Plenoptic Odometry", Zeller *et al.* present a Direct Plenoptic Odometry (DPO) algorithm for plenoptic cameras, which enables high quality camera tracking and environment mapping, leading to the reconstruction of semi-dense point clouds of large scenes. Volumetric reconstruction of transparent scenes such as flames is targeted in "Characterization of Plenoptic Imaging Systems and Efficient Volumetric Estimation from Plenoptic Data", by Anglin *et al.* The authors propose a novel refocusing transformation that makes the volumetric deconvolution shift-invariant, so that deconvolution is a viable solution that can be solved more efficiently. Guo *et al.* tackle the 3D reconstruction of small transparent and translucent specimen such as zebrafish using light microscopy in "A Two-phase 3D Reconstruction Approach for Light Microscopy Axial-view Imaging". They develop a multiview microscopy and propose a two-phase 3D reconstruction approach for accurate boundary reconstruction of these challenging biological samples. Reconstruction of surface Bidirectional Reflectance Distribution Functions (BRDF) is investigated by Lu *et al.* in "Identifying Surface BRDF from a Single 4D Light Field Image via Deep Neural Network", where a convolutional neural network (CNN) is proposed for BRDF identification from a single 4D light field image.

Due to the the high dimensional data and the limited sensor resolution, there is a problematic trade-off in the spatial and angular dimensions. The next three papers address super-resolution and reconstruction of a sparsely-sampled light field in the spatial and/or angular dimensions. In "Super Resolution of Light Field Images using Linear Subspace Projection of

Digital Object Identifier 10.1109/JSTSP.2017.2759458

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Patch-Volumes", Farrugia *et al.* present an example-based light field spatial super-resolution. They train linear projections between subspaces of reduced dimension in which patch-volumes extracted from the light field reside. They further extend their method for angular super-resolution. Instead of training the projections directly, Yang *et al.* use a manifold constraint in their paper "Image Reconstruction via Manifold Constrained Convolutional Sparse Coding for Image Sets". They take a Laplacian manifold constraint as the regularization term on the coding and dictionary, achieving consistent results. Different from the above learning-based approaches, Vagharshakyan *et al.* exploit sparse regularization of EPIs in the shearlet transform domain, in their paper "Accelerated Shearlet-Domain Light Field Reconstruction". They consider angular super-resolution as an inpainting problem on the EPI, and use an adapted discrete shearlet transform to super-resolve such angular resolution. The algorithm is implemented on the GPU for efficiency.

The next five papers address compression techniques. The first paper, "Comparison and Evaluation of Light Field Image Coding Approaches" by Viola *et al.* evaluates the perceived visual quality of light field images, and compares the performance of several state-of-the-art light field compression algorithms. The paper provides a useful benchmark for future development of compression solutions for light fields. In the second paper, "Pseudo Sequence based 2-D Hierarchical Coding Structure for Light-field Image Compression", Li *et al.* optimize the organization of the different views in a light field with an efficient coding order and a flexible reference structure, and propose a novel pseudo sequence based on a 2D hierarchical coding structure. Besides inter-frame prediction, intra-block prediction is also crucial for compression performance. Monteiro *et al.* propose in "Light Field Image Coding using High Order Intra Block Prediction" a novel, high order, intra-block prediction strategy, with modes ranging from pure translations to projective or bilinear transformations, optimized according to an appropriate rate-distortion criterion. Similarly, the fourth paper "Light Field Compression with Homography-based Low Rank Approximation" by Jiang *et al.* proposes the use of low-rank approximations to deal with different homography warping between views. The light field views are aligned using either one global homography, or multiple homographies depending on how much the disparity across views varies from one depth plane to the other. Last, lossless light field image compression is investigated in the fifth paper, "Minimum Description Length Sparse Modeling and Region Merging for Lossless Plenoptic Image Compression" by Helin *et al.*

Editing light fields is a non-trivial generalization of traditional 2D image editing, offering new possibilities but also challenges because of the richer information and higher dimensionality of a light field. The next four papers investigate new editing tasks from very different angles. In the first paper, Ortin *et al.* work on the problem of "Analyzing Interfaces and Workflows for Light Field Editing", and provide a detailed description of users' performance and preferences for a number of simple editing tasks (recoloring, material editing, contrast enhancement, etc.), which can benefit other researchers and designers when creating new light field editing tools and interfaces. The second paper "Image Reshaping for Efficient Compression of Plenoptic Content" by Jin *et al.* proposes a novel light field image reshaping scheme, which efficiently exploits the correlation among macropixels when compressing the lenslet images, using current block-based coding standards. In the paper entitled "Super-rays for Efficient Light Field Processing", Hog *et al.* focus on the problem of light field over-segmentation. Similar to super-pixels in a 2D image, they introduce the concept of the super-ray, which is defined as a group of rays within and across views in a light field. The fourth paper, "Design of 3D Cellular Automata-based Video Authentication with an Optical Integral Imaging Generated Memory-distributed Watermark" by Li *et al.* proposes the editing of watermarks into image and videos based on integral imaging techniques.

The final part of this special issue considers light field display problems. Currently, there are no methods nor practical tools to quantify a light field display's effective resolution, or the perceived quality of the presented imagery. The first paper entitled "Quantifying spatial and angular resolution of light field 3D displays" by Kovcs *et al.* derives the objective spatial resolution limit by using frequency domain analysis of recorded test patterns, and computes the angular resolution limit by measuring the display's ability to show uniform, angle-dependent patterns. Finally, in their paper "Layered Display with Accommodation Cue using Scattering Polarizers", Moon *et al.* investigate the long-term visual fatigue challenge in 3D display, i.e., the vergence-accommodation conflict (VAC). They propose a new type of compressive light field display which supports accommodation, by employing two stacked scattering polarizers to diffuse light rays, so that two-layer images can be displayed at different depths to mitigate the VAC.

It was our pleasure to put together this issue, and we hope you enjoy reading it as well!

Sincerely, The Guest Editors

YEBIN LIU, *Lead Guest Editor*
Tsinghua University
Beijing 100084, China

LU FANG, *Guest Editor*
Gradate School at Shenzhen
Tsinghua University
Shenzhen 518055, China

DIEGO GUTIERREZ, *Guest Editor*
Universidad de Zaragoza
Zaragoza 50009, Spain

QING WANG, *Guest Editor*
Northwestern Polytechnical University
Xian 710072, China

JINGYI YU, *Guest Editor*
ShanghaiTech University
Shanghai 201210, China
University of Delaware
Newark, DE 19716 USA

FENG WU, *Guest Editor*
University of Science and Technology of China
Hefei 230026, China

Yebin Liu received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2002, and the Ph.D. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2009. He has been working as a Research Fellow in the Computer Graphics Group, Max Planck Institute for Informatik, Saarbrücken, Germany, in 2010. He is currently an Associate Professor in Tsinghua University. His research areas include computer vision and computer graphics. He received Second Prize for the National Technical Invention Award in 2008, First Prize for the National Technical Invention Award in 2012, the Young Academic Research Fellow Award of Tsinghua University in 2013, and the NSFC Excellent Yong Research Fellow in 2015.

Lu Fang (SM'16) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2007, and the Ph.D. degree in electronic and computer engineering from Hong Kong University of Science and Technology, Hong Kong, in 2011. She is currently an Associate Professor in Tsinghua University, Beijing, China. She has published more than 60 peer-reviewed journal and conference papers, including the IEEE SPM, JSTSP, TIP, ICCV, CVPR, etc. Her research interests include computational photography and visual computing.

Diego Gutierrez is a Professor in the Universidad de Zaragoza, Zaragoza, Spain, where he leads the Graphics and Imaging Lab. He has published his research on physically based global illumination, computational imaging, and perception in top journals and conferences (including SIGGRAPH and Eurographics). He is currently the Papers Chair for Eurographics 2018, and EGSR 2012 and APGV 2011 before that. He has served on many Program Committees, including both SIGGRAPH Conferences and Eurographics. He is a Co-Editor-in-Chief of *ACM Transactions on Applied Perception*, and an Associate Editor of four other journals. He received the Google Faculty Research Award in 2014, and an ERC Consolidator Grant in 2016.

Qing Wang (M'05) received the B.S. degree in information mathematics from the Department of Mathematics, Peking University, Beijing, China, in 1991, and the Master's and Ph.D. degrees from the Department of Computer Science and Engineering, Northwestern Polytechnical University, Xian, China, in 1997 and 2000, respectively. He is currently a Full Professor in the School of Computer Science, Northwestern Polytechnical University. In 2006, he was a recipient of outstanding talent program of new century by the Ministry of Education, China. He is the author of more than 100 papers in the international journals and conferences. His research interests include computer vision and computational photography, such as three-dimensional reconstruction, light field imaging and processing, object detection, tracking, and recognition.

Jingyi Yu received the B.S. degree from California Institute of Technology, Pasadena, CA, USA, in 2000, and the Ph.D. degree from Massachusetts Institute of Technology, Cambridge, MA, USA, in 2005. He is the Director of Virtual Reality and Visual Computing Center, School of Information Science and Technology, ShanghaiTech University, Shanghai, China. He is also a Full Professor in the University of Delaware, Newark, DE, USA. He has published more than 100 papers at highly refereed conferences and journals including more than 50 papers at the premiere conferences and journals CVPR/ICCV/ECCV/TPAMI. He is currently an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Elsevier CVIU*, *Springer TVCJ,* and *Springer MVA*.

Feng Wu (M'99–SM'06–F'13) received the B.S. degree in electrical engineering from Xidian University, Xian, China, in 1992, and the M.S. and Ph.D. degrees in computer science from Harbin Institute of Technology, Harbin, China, in 1996 and 1999, respectively. He is currently a Professor in the University of Science and Technology of China, Hefei, China, and the Dean of the School of Information Science and Technology. Before that, he was the Principle Researcher and the Research Manager with Microsoft Research Asia. He is the author or coauthor of more than 200 high quality papers (including several dozens of the IEEE Transaction papers) and top conference papers on MOBICOM, SIGIR, CVPR, and ACM MM. He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEM FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, and several other International journals. He received the IEEE Circuits and Systems Society 2012 Best Associate Editor Award.

Light Field Image Processing: An Overview

Gaochang Wu, Belen Masia, Adrian Jarabo, Yuchen Zhang, Liangyong Wang, Qionghai Dai, *Senior Member, IEEE*, Tianyou Chai, *Fellow, IEEE*, and Yebin Liu, *Member, IEEE*

***Abstract*—Light field imaging has emerged as a technology allowing to capture richer visual information from our world. As opposed to traditional photography, which captures a 2D projection of the light in the scene integrating the angular domain, light fields collect radiance from rays in all directions, demultiplexing the angular information lost in conventional photography. On the one hand, this higher dimensional representation of visual data offers powerful capabilities for scene understanding, and substantially improves the performance of traditional computer vision problems such as depth sensing, post-capture refocusing, segmentation, video stabilization, material classification, etc. On the other hand, the high-dimensionality of light fields also brings up new challenges in terms of data capture, data compression, content editing, and display. Taking these two elements together, research in light field image processing has become increasingly popular in the computer vision, computer graphics, and signal processing communities. In this paper, we present a comprehensive overview and discussion of research in this field over the past 20 years. We focus on all aspects of light field image processing, including basic light field representation and theory, acquisition, super-resolution, depth estimation, compression, editing, processing algorithms for light field display, and computer vision applications of light field data.**

***Index Terms*—Light field imaging.**

I. Introduction

As THE most important medium for people to perceive the world, light rays carry abundant information of our 3D environment. Unlike conventional images, which record the 2D projection of the light rays by angularly integrating the rays at each pixel, a light field describes the distribution of light rays in free space. This allows to capture richer information from our world. An early model that describes the distribution of light rays was first defined by Gershun[1] in 1936 and further completed by Adelson and Bergen [2] in 1991. This model is known as the plenoptic function, and describes light in a scene as a function of position, angle, wavelength and time.

However, obtaining the full plenoptic function for a scene is challenging, due to the high dimensionality of the data. Alternatively, modern light fields [3] mainly focus on the distribution of light rays as a function of position and angle. With this lower dimensionality and the progress in both hardware and software, today's light field capture devices are portable and commercially available [4], [5], and can even be integrated in cell phones [6]. Moreover, they have been integrated in microscopy [7], allowing rapid scan-less volumetric photographs of biological specimens. The widespread of light field cameras has allowed several new applications, ranging from their initial purpose (photorealistic image-based rendering [3]), to current computer vision applications that make use of their rich encoded information; these include 3D reconstruction, segmentation and matting, saliency detection, object detection and recognition, tracking, and video stabilization.

Despite this large advance in the last 30 years, many scientific challenges remain for researchers to address. From an acquisition point of view, light field imaging requires the ability to capture higher dimensional data as opposed to simply recording a 2D projection in conventional photography (i.e., a light field camera should be designed to capture the directional light distribution at each location on the sensor). However, acquiring high-dimensional data typically imposes a resolution trade-off between the dimensions. Moreover, processing and analysis techniques for light fields are affected by their high dimensionality, which increases the computational complexity and imposes more challenging conditions on the design of algorithms. For example, segmentation in conventional 2D images aims to separate foreground and background within a single image, whereas when performing segmentation in light fields the photo-consistency must be kept.

This paper aims to provide an overview on the field of light field imaging and processing, while simultaneously revealing the challenges within. We first provide the theoretical description of light fields (Section II). Then, we dive into different light field processing areas, which can be roughly divided in three modules (see Fig. 1): *low-level hardware-based acquisition*, which focuses on the trade-offs needed to effectively capture a light field (Section III); *mid-level processing*, including dif-

Manuscript received January 7, 2017; revised May 22, 2017 and July 27, 2017; accepted August 3, 2017. Date of publication August 30, 2017; date of current version October 23, 2017. This work was supported in part by the National Key Foundation for Exploring Scientific Instrument 2013YQ140517, in part by the National NSF of China under Grants 61522111 and 61531014, in part by the Spanish Ministry of Economy and Competitiveness (TIN2016-79710-P), in part by the European Research Council (Consolidator Grant, project Chameleon, ref. 682080), and in part by Science and Technology Planning Project of Guangdong Province, China (2015B010105004). The guest editor coordinating the review of this paper and approving it for publication was Dr. Lu Fang. *(Corresponding author: Yebin Liu.)*

G. Wu is with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110004, China, and also with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: 790723977@qq.com).

B. Masia and A. Jarabo are with the Graphics and Imaging Lab, Universidad de Zaragoza, Zaragoza 50009, Spain (e-mail: bmasia@unizar.es; adrianjarabo@gmail.com).

L. Wang and T. Chai are with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110004, China (e-mail: wangliangyong@mail.neu.edu.cn; tychai@mail.neu.edu.cn).

Y. Zhang is with the Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: slyzzhangyc@163.com).

Q. Dai and Y. Liu are with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: qionghaidai@mail.tsinghua.edu.cn; liuyebin@mail.tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2747126

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

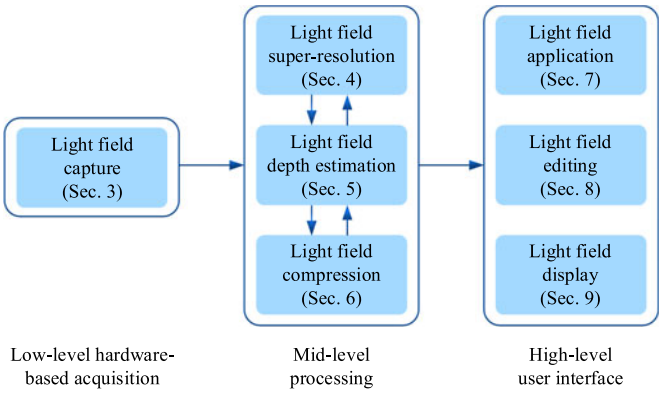


Fig. 1. Organizational framework of the paper.

ferent techniques proposed to mitigate the resolution trade-off (Section IV), depth estimation (Section V), and light field evaluation approaches and compression schemes (Section VI); and *high-level user interface or application-related algorithms*, including light field applications in vision and graphics (Section VII), editing techniques (Section VIII) and algorithms for display (Section IX). We additionally provide an overview of different databases and code publicly available for research (Section X).

The classification of light field research modules is not unique, and we adopt the classification above based on the following considerations. First, depth estimation is classified as a mid-level processing technique because depth information is the most important intermediate result for the deploy of editing and display algorithms, as well as computer vision applications. Second, light field super-resolution (or reconstruction) takes place after light field acquisition, and is often iteratively optimized with depth estimation; therefore, we classified it as a mid-level processing technique as well. Third, light field compression techniques are well related with light field super-resolution (reconstruction) approaches, and are thus also classified as middle-level processing techniques. Last, light field editing and display are two visualization problems, which belong to applications of computer graphics, while the applications we described in the paper are mainly within the field of computer vision.

Partially overlapping works can be found at an early introductory-level overview [8], at a more conceptual and theoretical approach [9], and in a broad survey on computational plenoptic imaging techniques [10]. The difference in this overview is that we focus on the 4D light field as a scene representation structure, and provide comprehensive overviews and discussions of all the possible processing techniques related with it, i.e., from the low-level hardware-based acquisition, through mid-level data processing, and all the way to the so-called high-level techniques related to visualization or editing.

II. Light Field Theory

In this section, we present a brief overview of light field theoretical principles. This includes the representation of a light field as a multi-dimensional function, the basics of light field visualization and rendering, and an analysis of light field sampling in the Fourier domain.

A. Light Field Representation

The plenoptic function [2] is a multidimensional function describing the set of light rays traveling in every direction through every point in 3D space, from a geometric optics perspective. To obtain such a function, one must measure the light rays at every possible location (x, y, z), from every possible angle (θ, ϕ), at every wavelength γ and at every time t. The plenoptic function is then a 7D function denoted as $L(x, y, z, \theta, \phi, \gamma, t)$.

However, such high dimensional data is difficult to record and handle in practice. Thus, the light field model has been simplified twice for practical usage. In the first simplification, the measured function is assumed to be monochromatic and time-invariant. The wavelength γ of each light ray is recorded independently in different color channels. And the time sequence t can be recorded in different frames for a dynamic light field (i.e., a light field video). In this way, we can remove the wavelength γ and time t dimensions from the plenoptic function, reducing the model from seven to five dimensions.

The second simplification was made by Levoy and Hanrahan [3] and Gortler *et al.* [11], who realized that the 5D representation still contained some redundancy and could be reduced to 4D by assuming that the light field was measured in free space. In such cases, light ray radiance remains constant along a straight line, making one dimension redundant in the 5D plenoptic function. The extra dimension increases the complexity of measuring and reconstructing the plenoptic function.

When parameterizing a 4D light field, there are three key issues [3]: computational efficiency, control over the set of rays, and uniform sampling of the light field space. Based on these issues, the most common solution to the representation of a 4D light field is to parameterize the light rays by the coordinates of their intersections with two planes placed at arbitrary positions. The coordinate system is denoted by (u, v) for the first plane and (s, t) for the second plane. An oriented light ray defined in the system first intersects the uv plane at coordinate (u, v) and then intersects the st plane at coordinate (s, t), and is thus denoted by $L(u, v, s, t)$ (the schematic diagram will be presented in Section II-C). Thus, the plenoptic function describing a light field is reduced from 7 to 4 dimensions, and parameterized by four coordinates (u, v, s, t).

B. Light Field Visualization

Although the function $L(u, v, s, t)$ is a simplified light field model, it is still hard to imagine this 4D representation. In this subsection, we will visualize this 4D light field in two different ways: integral light field structure, and 2D slices.

In the two-plane light field model $L(u, v, s, t)$, we can consider the st plane as a set of cameras with their focal plane on the uv plane. Two different perspectives can be taken to understand this model. First, that each camera collects the light rays leaving the uv plane and arriving at a point on the st plane (the collection of light rays from a certain viewpoint). Thus the 4D light field can be represented as a 2D array of images, such as the one

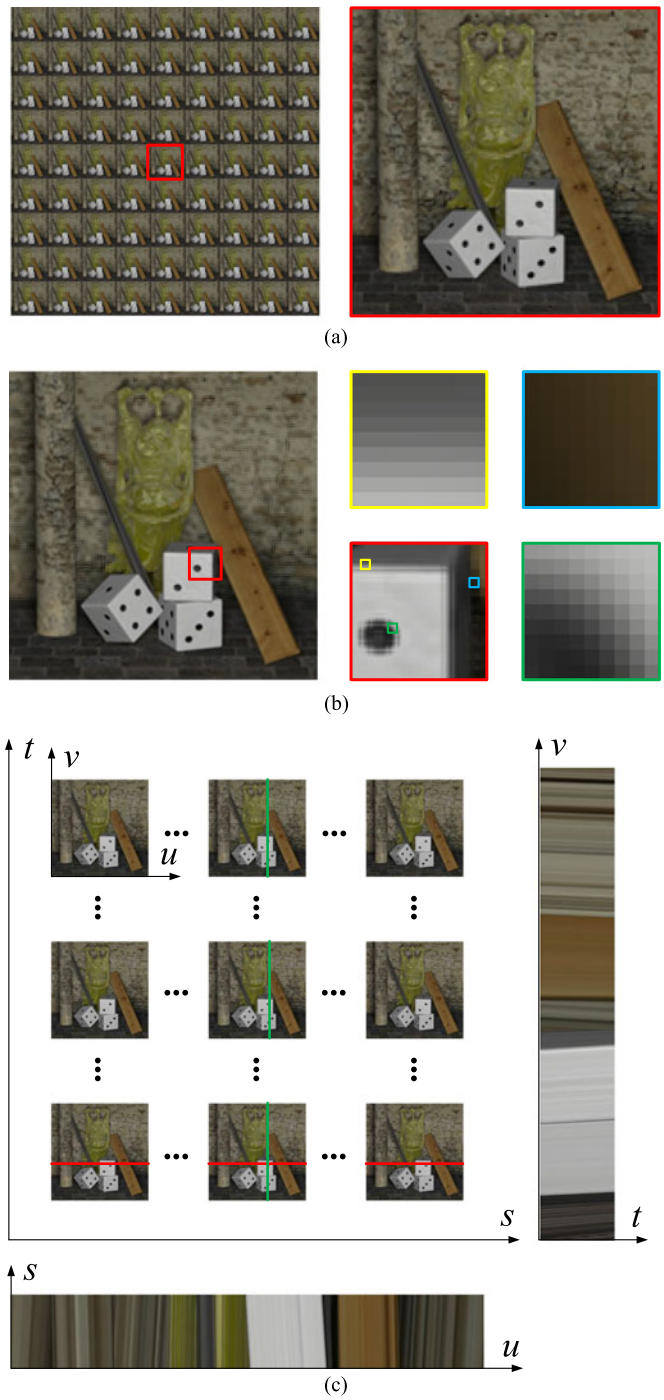


Fig. 2. Different ways of visualizing a light field. (a) A sub-aperture image $I_{s^*,t^*}(u,v)$ can be acquired by gathering the light field samples with fixed st coordinates (s^* and t^*). (b) A light field subview $I_{u^*,v^*}(s,t)$ can be acquired by gathering the samples with fixed uv coordinates (u^* and v^*). (c) The epipolar-plane image (EPI) is obtained by fixing the coordinates in both a spatial and an angular dimension. For example, the bottom EPI $E_{v^*,t^*}(u,s)$ is obtained by fixing v and t, and the right EPI $E_{u^*,s^*}(v,t)$ is obtained by fixing u and s.

shown in Fig. 2(a). Each image recorded by the camera is called a sub-aperture image, also known as a pinhole view. Second, that a certain point on the uv plane represents the light rays bound for all points on the st plane (the collection of light rays from different viewpoints projected onto a certain point, i.e., the same point as seen from different viewpoints, see Fig. 2(b)). Because the number of samples in the st plane depends on the number of viewpoints, while that in the uv plane depends on the camera resolution, the s and t dimensions are referred to as the angular dimensions and u and v dimensions are referred to as the spatial dimensions.

Specifically, for a 4D light field $L(u,v,s,t)$, the 2D slice $I_{s^*,t^*}(u,v)$ can be acquired by gathering the samples at the fixed st coordinates s^* and t^*. The slice $I_{s^*,t^*}(u,v)$ (sub-aperture image) can be considered as a photograph captured by a camera located at (s^*,t^*). The right part in Fig. 2(a) visualizes a sample sub-aperture image. Similarly, the slice $I_{u^*,v^*}(s,t)$ is acquired by gathering the samples at the fixed uv coordinates u^* and v^*. The slice $I_{u^*,v^*}(s,t)$ (often referred to as the light field subview) is formed by gathering, at each point, the rays from different viewpoints, as visualized in Fig. 2(b) for three points (marked in yellow, green and blue).

The 2D slices described above are acquired by gathering either two spatial dimensions or two angular dimensions. By gathering the light field samples with a fixed spatial coordinate v and an angular coordinate t (or u and s), one can produce a slice $E_{v^*,t^*}(u,s)$ (or $E_{u^*,s^*}(v,t)$). The resulting map is called the epipolar-plane image (EPI), which is a well-known term from multi-view computer vision [12]. Unlike a pinhole image or a light field subview, the EPI contains information in both spatial and angular dimensions. Consider a certain point on the Lambertian surface of an observed object with depth Z; when changing the coordinate s, the coordinate u will also change according to $\Delta u = \frac{f}{Z}\Delta s$ (where f is the distance between the planes), forming a line on the EPI. Points with different depths can be visualized as lines with different slopes in the EPI, see Fig. 2(c). Conversely, the slopes of lines in the EPI reflect the depth of the scene captured by the light field. This particular structure is widely exploited to infer scene geometry, as will be introduced in Section V.

C. Light Field Rendering

We now know that a 4D light field $L(u,v,s,t)$ can be considered as a set of views captured by cameras defined by the two parallel planes, st and uv. In practice, the st and uv space can also be spherical. Compared with planar planes, spherical planes sample light rays more uniformly in directional space and provide a fixed distance to the object. Fig. 3 illustrates a light field sampling system with spherical planes [13]. An object is located at the very center of a sphere with cameras distributed on its surface. The cameras capture a 4D light field of the object as long as the sphere is larger than its convex hull.

With a sufficiently dense set of cameras, one can create virtual renderings of the light field at any position of the sphere surface—or even closer to the object—by resampling and interpolating the light rays [3], rather than synthesizing the view based on geometry information [14]. Fig. 3(a) shows a virtual view (located at the blue dot) that is closer to the object (the viridis shape) than the existing views. Because the radiance of the rays remains constant in free space, some of the light rays can be resampled from the existing views. However, there are

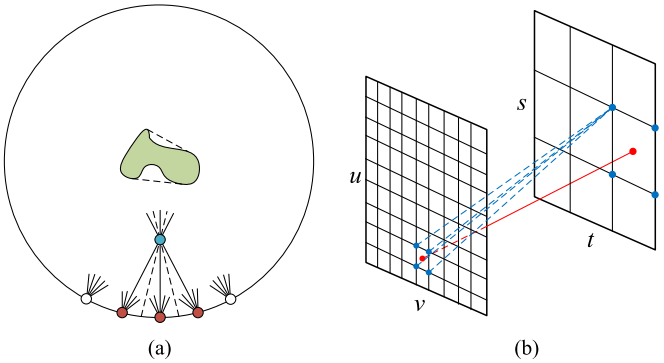


Fig. 3. Rendering a virtual view using the recorded light field. (a) For a virtual view located at the blue dot that is closer to the object but outside its convex hull (the dashed line around the center object), some rays can be resampled in the existing views (e.g., the rays in the solid lines can be resampled by the rays captured by the cameras that are located at the red dots. However, some rays are captured by none of the cameras, such as the rays in the dashed lines. (b) For a virtual ray, the nearest 16 sampled rays are used to perform an interpolation (only four are depicted in the figure for illustration purposes).

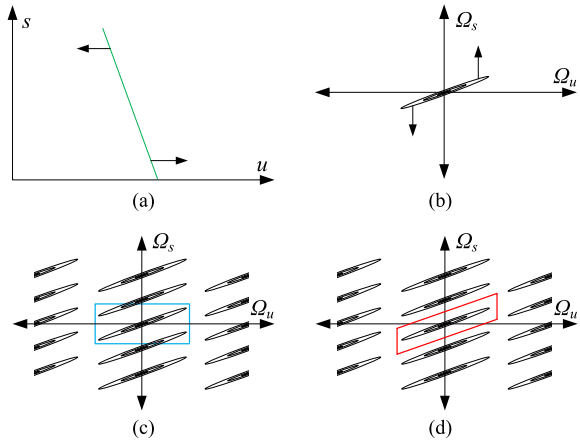


Fig. 4. (a) An EPI with a single line located slightly off the u-axis. (b) The power spectrum of the EPI in (a). The slope of the power spectrum with Ω_u-axis is determined by the slope of the line with the u-axis in the EPI. (c) Sampling the light field produces copies of the power spectrum in the Fourier domain. Using the reconstruction filter shown by the blue box will cause high frequency leakage from the neighboring copies of the power spectrum, leading to aliasing in the form of ghosting. (d) Using an alternative reconstruction filter (shown by the red box), we can accurately reconstruct the original EPI in (a).

still light rays that none of the cameras capture, and that must therefore be inferred through interpolation (e.g., the rays shown by dashed lines in Fig. 3(a)). To render such rays, first, their intersection coordinates with the two planes are computed. Then, the nearest 16 sampled rays are used to interpolate the virtual ray. Fig. 3(b) depicts the two-plane representation of the 4D light field introduced in Section II-A, in which four of the sampled rays are applied for interpolation. The above idea is called light field rendering [3].

In light field rendering, insufficient samples will cause ghosting effects in the novel views. However, it is impractical to acquire too many samples of a light field. For example, the synthetic *hallway* light field rendered by Levoy and Hanrahan [3] uses four light slabs, each of which involves 64×32 images with a raw size of 1608 MB. Chai *et al.* [15] and Lin *et al.* [16] investigated the minimum number of samples needed for light field rendering and concluded that the contributing pixels must at least touch each other to render novel views without ghosting. In other words, the maximum disparity between the neighboring views must be less than 1 pixel, a value closely related to the camera resolution and scene depth; the closer together the contributing pixels are (i.e., the less disparity), the sharper the interpolated point will be.

When the geometric information of the scene is known, the required number of samples can be greatly reduced, and even an irregularly sampled light field is appropriate for novel view rendering. Buehler *et al.* [17] presented a generalized model that combines light field rendering and depth image-based rendering using irregularly sampled images (i.e., unstructured input). To render a novel view, the input views were combined using a "camera blending field", that was related to the angular difference, estimates of undersampling, and the field of view (FOV).

D. Fourier Analysis

In this subsection, we focus on the analysis of light field sampling in the Fourier domain. Consider an EPI with a single line located slightly off the u-axis as shown in Fig. 4(a). The power spectrum of the EPI is shown in Fig. 4(b), where the intersection angle with the Ω_u-axis is determined by the slope of the line in the EPI. Sampling the light field produces copies of the spectrum shown in Fig. 4(c). For a light field with sufficiently dense sampling in the angular dimensions, there will be no overlap between the copies along the Ω_s axis in the Fourier domain. Using the reconstruction filter shown by the blue box in Fig. 4(c) will cause high frequency leakage from the neighboring copies of the power spectrum, producing ghosting effects in the reconstructed views [36]. An alternative is to use a shear reconstruction filter (shown by the red box in Fig. 4(d)). In this way, the EPI shown in Fig. 4(a) can be accurately reconstructed.

However, issues in practical cases are usually more complicated. Fig. 5(a) shows an EPI containing three lines with different slopes intersecting each other. The resulting power spectrum of the sparsely sampled light field is shown in Fig. 5(c), where the neighboring copies overlap in the high frequency spectrum. Using the shear reconstruction filter shown by the box with dashed lines in Fig. 5(c) will still cause high frequency leakage from other copies of the power spectrum, producing ghosting effects. To reconstruct the light field without those ghosting effects, the high frequencies should be filtered properly (shown in Fig. 5(d)), producing a band-limited reconstructed light field [37]. The resulting power spectrum after filtering (shown in Fig. 5(e)) is close to the one before the sparse sampling. The reconstructed EPI shown in Fig. 5(f) is slightly blurrier than the original EPI due to the high frequency loss.

In addition, Fourier analysis of light fields has been extensively used in light transport [38], [39] and wave optics [40]–[42]. Ren [43] derived a 4D "Fourier slice photography theorem", and indicated that a photograph formed with a full lens aperture is a 2D slice in the 4D light field. This mathematical theory was applied to analyze the performance of digital refocusing, and further used to yield a Fourier-domain approach for digital refocusing using light fields. Zhang and Levoy [41]

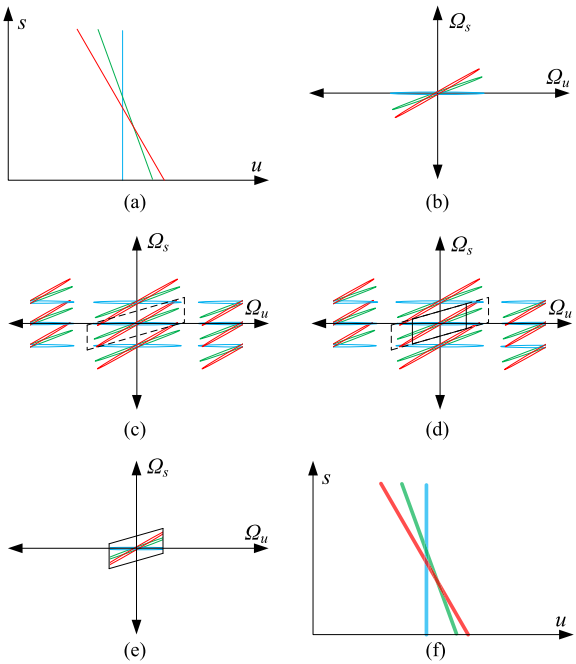


Fig. 5. (a) An EPI with three intersecting lines. (b) The power spectrum of the EPI in (a). (c) In the power spectrum of the sparsely sampled light field, the neighboring copies are overlap in the high frequency portion. Using the shear reconstruction filter shown by the box with dashed lines will still cause high frequency leakage. (d) The high frequency portion is filtered properly, producing the reconstruction filter shown by the box in the solid line. The filtered power spectrum is shown in (e), and the reconstructed EPI is shown in (f).

indicated that the light field is equivalent to a smoothed Wigner distribution, which approaches the Wigner distribution under a geometric optics approximation. They further explored the correspondence between Fourier slice photography and wavefront coding using the Wigner distribution and light fields.

III. Light Field Acquisition

This section focuses on existing devices or methods for light field acquisition. A conventional camera captures a 2D projection of a light field on its sensor plane by integrating the light rays that hit each pixel from every direction. In contrast, the devices or methods for light field acquisition measure the distribution of light rays in a directionally-resolved manner, avoiding angular integration.

However, sensors can only measure the information from two dimensions (usually two spatial dimensions) of a scene at a single moment. To acquire a 4D light field, we need to capture multiple samples along the angular dimensions. Existing light field acquisition approaches can be divided into three fundamental categories [8]–[10], [44]: multi-sensor capture, time-sequential capture and multiplexed imaging. In the following, we will provide a detailed review of the existing light field acquisition approaches according to those three categories.

A. Multi-Sensor Capture

The multi-sensor capture approach requires an array of image sensors distributed on a planar or spherical surface to simultaneously capture light field samples from different viewpoints

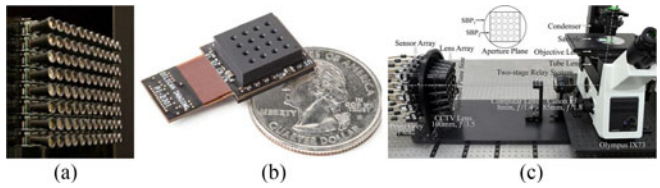


Fig. 6. Devices for multi-sensor capture. (a) A camera array system composed of 8×12 video cameras developed by Wilburn *et al.* [20]. (b) The Picam developed by Venkataraman *et al.* [22] was an ultra-thin monolithic 4×4 camera array that could be integrated into a cell phone. (c) Lin *et al.* [23] used a 5×5 camera array system to capture the light field of microscopic objects.

[3]. The spatial dimensions (u and v) of the light field are determined by the sensors, while the angular dimensions (s and t) are determined by the number of cameras and their distribution. Therefore, the 4D light field is recorded by the combination of the captured images.

In 2002, Yang *et al.* [18] described a design using an array of 8×8 video cameras for dynamic light field capture. To overcome data bandwidth problems, they employed a distributed rendering algorithm. In the same year, Wilburn *et al.* [45] applied 6 CMOS image sensors to record a synchronized video dataset. Each camera had a processing board that implemented MPEG2 compression to achieve a scalable capture. They further expanded the system to 125 video cameras and used it to capture multi-thousand frame-per-second (fps) videos [20], [46] (one configuration of this system is shown in Fig. 6(a)). Zhang and Chen [19] presented a system consisting of 6×8 cameras able to render novel views and reconfigure the camera positions to achieve better rendering quality. Chan *et al.* [47] used a 1×8 array of video cameras to capture and render a dynamic image-based representation called "plenoptic video". The system was able to stream plenoptic video (256×256 spatial resolution) at 15 fps over the network. Liu *et al.* [48] also developed an 8×8 dynamic light field streaming system over broadband network.

Typical camera array systems are bulky and expensive and are thus unsuitable for most commercial uses. The ProFusion 25 [21], manufactured by ViewPLUS Inc., includes an array of 5×5 VGA cameras packaged inside a small box and is able to capture light field video at 25 fps. Venkataraman *et al.* [22] presented an ultra-thin monolithic camera array, which they called PiCam (Pelican Imaging Camera-Array, Fig. 6(b)). The PiCam is a 4×4 camera array; each camera is able to capture 1000×750 pixels. The entire device is smaller than a coin, and can be integrated into a cell phone [6].

Moving to smaller scene scales, Lin *et al.* [23] presented a camera array-based light field microscopy setup to measure the distribution of light passing through a specimen volume. The system, shown in Fig 6(c), comprised a 5×5 sensor array operating at a resolution of 1024×768 at 30 fps.

The multi-sensor capture approach is able to capture a light field instantaneously, and is thus competent to record a light field sequence. Early multi-sensor systems were inevitably bulky and expensive due to large amounts of video cameras. However, we can see the potential of this light field acquisition method through the recent cheaper and more portable designs.

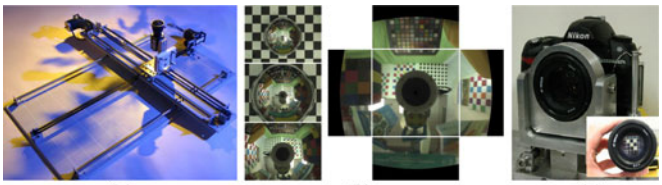


Fig. 7. Devices for time-sequential capture. (a) A motorized linear stage for mounting a single camera to capture 4D light fields by Unger *et al.* [49]. (b) The *axial light field* proposed by Taguchi *et al.* [27], consisting of a mirror ball and a single camera. The light field was captured by moving the camera along the mirror's axis of rotation. (c) A programmable aperture presented by Liang *et al.* [26] was applied to capture light fields through multiple exposures.

B. Time-Sequential Capture

In contrast with the multi-sensor approach, a time-sequential capture approach uses a single image sensor to capture multiple samples of the light field through multiple exposures. The typical approach uses a sensor mounted on a mechanical gantry to measure the light field at different positions [3], [11].

Two gantry systems were presented by the Computer Graphics Laboratory at Stanford University [24]: one was a computer-controlled gantry with four degrees of freedom, translation in X and Y, nod and shake; another was a *Lego Mindstorms* gantry in which the motors have rotatory controllers enabling the camera to move along accurate and well-defined paths. Unger *et al.* [49] used a single camera mounted on a motorized linear stage with two degrees of freedom, translation in X and Y, to capture light fields. Fig. 7(a) shows the system by Unger *et al.* [49]. A similar design with only one degree of freedom was presented by Kim *et al.* [25]. The system was applied to capture 3D outdoor light fields with high angular resolution. To acquire a light field with large FOV, Dansereau *et al.* [50] presented a compact optical design combining a monocentric lens with microlens arrays (which will be introduced in the following section). The system was mounted on a mechanical arm that can be rotated around a fixed axis to enable large FOV capture.

The approaches above capture images at different viewpoints by moving the image sensor, which requires high-precision control and is time consuming. Fortunately, some fast time-sequential capture approaches have also been introduced. Ihrke *et al.* [51] proposed using a planar mirror and a high dynamic range video camera to record light fields. The moving mirror produces different viewpoints to be captured by the camera. Taguchi *et al.* [27] proposed a system consisting of a mirror ball and a single camera, as shown in Fig. 7(b). The light field, which they called "axial light field", was captured by moving the camera along the mirror's axis of rotation.

Rather than applying a reflective surface, Liang *et al.* [26] presented a programmable aperture for light field acquisition, shown in Fig. 7(c). The aperture patterns were encoded into an opaque slit of paper or a liquid crystal array, allowing the camera to capture the light rays from certain angles. By synchronizing the aperture patterns with the camera exposures, the light field can be acquired at full sensor resolution. The entire capturing procedure required 25 patterns for a light field with 5×5 angular resolution, and required 10 to 20 ms for each exposure. The spatial resolution of each image was 3039×2014, but was downsampled to 640×426 to achieve high computational efficiency.

Compared with multi-sensor systems, time-sequential capture systems require only a single sensor, which reduces the cost of the entire system. In addition, time-sequential capture systems are able to capture light fields at dense angular resolutions, which multi-sensor systems cannot do because of the high associated cost. However, the capture process in time-sequential systems is always time-consuming; thus, they are appropriate only for static scenes.

C. Multiplexed Imaging

The last approach aims to encode the 4D light field into a 2D sensor plane, by multiplexing the angular domain into the spatial (or frequency) domain. It allows dynamic light field capture with a single image sensor, but imposes a trade-off between the spatial and angular resolutions (i.e., one can obtain densely sampled images in the spatial domain with sparse samples in the angular domain, and viceversa). Multiplexed imaging can be further divided into spatial multiplexing and frequency multiplexing.

1) Spatial Multiplexing: In spatial multiplexing, an interlaced array of elemental images representing samples from different 2D slices of the light field are captured by the sensor. Most spatial multiplexing approaches are implemented using an array of microlenses or a lenslet array mounted on the image sensor. Interestingly, this is one of the first approaches for light field imaging: in 1908, Lippmann used this approach in the development of his "integral photography" [52].

In 1992, Adelson and Wang [53] described a novel camera for "single lens stereo" that incorporates a main lens and a lenticular array placed before the sensor. The device could capture a 5×5 angular resolution light field with 100×100 pixels in each image. Ng *et al.* [28] presented a hand-held light field camera by inserting a 296×296 lenslet array between the sensor and main lens. This device looked and operated like a conventional camera and could capture 14×14 angular resolution light fields.

The spatial multiplexing devices described above belong to the type referred to as "plenoptic camera 1.0"; in it, each microlens captures one position in the scene. Two commercially available cameras that fall under the "plenoptic camera 1.0" category are those by Raytrix [5] and Lytro [4]. They are designed for industrial and consumer use, respectively; Fig. 8(a) shows the commercially available "Lytro Illum". Wei *et al.* [54] proposed a novel design that introduced sample irregularities and lens aberrations into a light field camera to improve its quality and usability. Each microlens in a "plenoptic camera 1.0" captures the angular distribution of the radiance. By gathering pixels in the same coordinate of each subview, we can obtain an image located at a certain viewpoint. An alternative version uses multiple microlenses to capture the same position in a scene, referred to as "plenoptic camera 2.0", and the angular information is spread across each microlens [55].

Instead of using a lenslet array, filters and mirror arrays are also employed in some light field capture approaches. Horstmeyer *et al.* [56] used multiple filters placed in the pupil



Fig. 8. Devices for multiplexed imaging. (a) Lytro Illum [4], a "plenoptic camera 1.0" that is currently commercially available. It uses a lenslet array to interlace the images of different views on a single sensor. (b) An integral camera that uses a 4×5 array of lenses and prisms to capture a low angular resolution light field developed by Georgiev *et al.* [30]. The angular resolution can be upsampled by view interpolation techniques. (c) Marwah *et al.* [33] described an attenuation mask-based scheme for frequency multiplexing capture. Different samples in the light field were encoded into different frequency bands.

plane of the main lens and a pinhole array on the sensor plane to capture light fields. The filters could be exchanged and reconfigured, providing a high degree of flexibility. Similarly, Manakov *et al.* [31] developed a single reconfigurable add-on that enables plenoptic imaging with standard cameras including multispectral, high dynamic range and light field imaging. Similar to the "axial light field" described in Section III-B that captured a light field reflected by a spherical mirror, Unger *et al.* [49] and Lanman *et al.* [29] presented an array of spherical mirrors to capture light fields. A slight difference is that their designs used multiple mirrors and thus turned the time-sequential capturing into spatial multiplexing.

In addition to measuring macroscopic objects or scenes, spatial multiplexing has also been applied to capture specimens at micro-scales [7], [57]–[59]. Levoy *et al.* [7], [57] inserted a microlens array into the optical path of a conventional microscope to record light fields of biological specimens on a single sensor plane. Compared with light field microscopy systems based on camera arrays [23] discussed in Section III-A, Levoy's system captures light fields with high angular resolution (17×17), but at reduced spatial resolution. Debise and Ihrke [60] developed a light field microscope using a consumer light field camera, Lytro, and an optical matching system, and achieved a maximum spatial resolution of about 6.25 μm.

Spatial multiplexing has been the most extensively used approach to capture light fields. It enables capturing a light field in a single image with a single exposure. However, there is a problem inherent to spatial multiplexing approaches, and that is the trade-off between angular and spatial resolution at the image sensor. To handle this problem, Georgiev *et al.* [30] sacrificed angular resolution for higher spatial resolution in light fields. They presented a 4×5 array of lenses and prisms cut into squares placed in front of a conventional camera (shown in Fig. 8(b)) to capture a low angular resolution light field. Then, interpolation techniques were applied to synthesize a dense angular resolution light field. Many researchers have focused on super-resolving a light field in the spatial and/or angular domain; please refer to Section IV for more details on super-resolution.

2) Frequency Multiplexing: Unlike the spatial multiplexing approach, which interlaces the 2D light field slices on the sensor plane, the frequency multiplexing approach encodes different 2D slices of the light field into different frequency bands. Typically, frequency multiplexing approaches use a modulation mask to achieve a certain property in the Fourier domain [44].

Veeraraghavan *et al.* [32] described a theoretical framework for 4D light field acquisition using an attenuation mask in the optical path of a conventional image sensor; they termed the approach "Dappled Photography". Rather than blending the light rays, the patterned mask attenuated and encoded them on the image sensor. The Fourier transformed image is rearranged into 4D planes and then an inverse Fourier transform is applied to restore the light field. As indicated by Lanman *et al.* [61], the patterned masks are equivalent to a truncated Fourier series approximation of a pinhole array for high angular sampling rates. Ashok and Neifeld [62] further described two separate architectures for light field imaging allowing compressive light field imaging in either the angular or spatial dimensions. In their designs, an amplitude mask was employed behind the main lens for angular compressive light field imaging; and an amplitude mask was placed over each microlens, for spatial compressive light field imaging. Antipa *et al.* [34] encoded multiplexed spatial and angular information using a diffuser plate. Compared with an attenuation mask, the diffuser plate allows higher light throughput, and provides an inexpensive and flexible means for frequency multiplexing. Pégard *et al.* [63] applied the compressive light field imaging technique to capture specimens at microscopic scales, while Cohen *et al.* [64] applied a wavefront coding technique that produces a more uniform distribution across depth to enhance the performance of the light field microscope.

Taking advantage of learning techniques, Marwah *et al.* [33] described an attenuation mask-based scheme and used an overcomplete dictionary to reconstruct the light field from its coded 2D projection (shown in Fig. 8(c)). Wang *et al.* [65] further improved the quality of reconstructed light fields using a random convolution CMOS sensor, which is able to maintain more information by means of correlation.

Although frequency and spatial multiplexing are two different multiplexed imaging approaches, they actually have a close relationship in terms of their multiplexed patterns. Ihrke *et al.* [44] indicated that patterns derived for Fourier space can be reconstructed by spatial interpolation, and spatial patterns can also be reconstructed with a Fourier space algorithm. In addition, Fourier reconstruction for frequency multiplexing is equivalent to spatial reconstruction with a sinc filter kernel, and the resulting ringing artifacts can be significantly removed by a spatial filter, such as cubic interpolation or an edge-preserving filter.

D. Discussion

We have introduced various light field acquisition approaches in this section, and Table I provides a summary of some of the typical ones. As we can see in the table, most multi-sensor approaches are able to capture light field video but require a large number of cameras. Meanwhile, time-sequential approaches use a single camera, and are able to capture high spatial and angular resolution light fields, but its time-consuming capturing procedure makes them impractical for dynamic scenes. Multiplexed imaging enables the capability of high-speed light field capture with a single camera, but leading to the problem of a

TABLE I
A SUMMARY OF TYPICAL LIGHT FIELD ACQUISITION APPROACHES

Category	Approach	Year	Implementation	Resolution	Capture speed	Notes
Multi-sensor	Yang *et al.* [18]	2002	8×8 camera array	$320 \times 240 \times 8 \times 8$	15–20 fps	–
	Zhang and Chen [19]	2004	6×8 camera array	$320 \times 240 \times 6 \times 8$	15–20 fps	–
	Wilburn *et al.* [20]	2005	10×10 camera array	$640 \times 480 \times 10 \times 10$	30 fps	–
	ProFusion 25 [21]	–	5×5 camera array	$640 \times 480 \times 5 \times 5$	25 fps	Portable device
	PiCam [22]	2013	4×4 camera array	$1000 \times 750 \times 4 \times 4$	–	Integrated unit
	Lin *et al.* [23]	2015	5×5 camera array	$1024 \times 768 \times 5 \times 5$	30 fps	Microscopy
Time-Sequential	Light field gantry [24]	2002	Gantry	$1300 \times 1030 \times 62 \times 56$	5 h/slab	–
	Kim *et al.* [25]	2013	Linear stage	$5616 \times 3744 \times 100$	–	3D light field
	Liang *et al.* [26]	2008	Programmable aperture	$3039 \times 2014 \times 5 \times 5$	above 0.5 s	–
	Taguchi *et al.* [27]	2010	Mirror ball	–	–	Axial light field
Multiplexed (spatial)	Ng *et al.* [28]	2005	Microlens array	$292 \times 292 \times 14 \times 14$	16 ms	Portable device
	Lanman *et al.* [29]	2006	Spherical mirror array	–	–	–
	Georgiev *et al.* [30]	2006	Lens-prism pairs	$700 \times 700 \times 4 \times 5$	–	–
	Manakov *et al.* [31]	2013	Camera add-on	$1800 \times 1200 \times 3 \times 3$		–
	Lytro Illum [4]	2014	Microlens array	$625 \times 434 \times 15 \times 15$	3 fps	Portable device
	Levoy *et al.* [7]	2006	Microlens array	$120 \times 120 \times 17 \times 17$	1/15 s	Microscopy
Multiplexed (frequency)	Veeraraghavan *et al.* [32]	2007	Attenuation mask	$228 \times 181 \times 9 \times 9$	–	–
	Marwah *et al.* [33]	2013	Attenuation mask	$480 \times 270 \times 5 \times 5$	–	Learning-based
	Antipa *et al.* [34]	2016	Diffuser plate	$170 \times 170 \times 11 \times 11$	–	–

Note that for multi-sensor and time-sequential capture, the spatial resolution refers to the sensor resolution; and for multiplexed imaging, the spatial resolution refers to the resolution of a sub-aperture image after being extracted by the corresponding algorithm described in each reference. The spatial resolution for the Lytro Illum [4] refers to that of a sub-aperture image extracted with the Light Field Matlab Toolbox [35].

resolution trade-off in the spatial and angular dimensions. Current light field acquisition devices at the commercial level use spatial multiplexed imaging. However, due to the large volume of data that needs to be recorded, it is hard to meet the capture speed needed for light field video capture; for instance, Lytro Illum [4] can only record light field video at 3fps. For frequency multiplexing, the modulation mask will inevitably lose luminous flux, leading to longer exposure times for each frame. It is worth mentioning that, as sensors get smaller and cheaper, multiple sensors can be assembled into a portable box (e.g., ProFusion 25 [21]), or even integrated onto a small monolithic camera array (e.g., PiCam [22]). Each sensor is able to record video at a high frame rate (e.g., ProFusion 25 [21] can record light field video at 25 fps). With this trend, light field cameras could find its way into mobile devices, such as mobile phones and tablet computers, in a near future.

IV. LIGHT FIELD SUPER-RESOLUTION

Due to their portability and low cost, plenoptic cameras have sparked renewed interest in light field imaging [4], [5]. However, limited by the sensor resolution and the processing speed, it is still very challenging to capture a high spatio-angular resolution light field and high frame-rate light field video. Moreover, the spectral range and resolution of current plenoptic cameras is in most cases restricted to that of three (RGB) channels.

To mitigate these problems on the basis of existing hardware, many researchers have focused on light field super-resolution in different dimensions. In this section, we investigate light field super-resolution mainly in terms of the spatial and angular dimensions, which have received the most attention. We will nevertheless, although more briefly, also look into approaches for super-resolving the spectral and temporal dimensions.

A. Spatial Super-Resolution

The light field captured by the Lytro Illum [4] is 7728×5368 pixels—higher than most current DLSR cameras; however, the resolution of a single sub-aperture image extracted by Light Field Matlab Toolbox [35] is only 625×434. A straightforward way to super-resolve the light field in the spatial dimensions is to employ single-image super-resolution techniques to upsample each of the views in the light field. However, this type of method treats each image in the light field as an individual entity, which of course results in underuse of the full potential of the information in the light field. As indicated in [55], [75], the true spatial resolution of a sub-aperture image is not limited by lenslet resolution, i.e., one can extract a sub-aperture image by taking advantage of the properties of a light field. In a 4D light field captured by plenoptic cameras, each image is a 2D sample from a viewpoint with only a slight difference with respect to its neighboring views, leading to non-integral shifts between two corresponding pixels in the images. By taking advantage of these non-integral shifts, pixels in the neighboring views can be propagated to the target view [55] (see Fig. 9).

Following the general principle described above, typical spatial super-resolution methods first estimate the depth of the scene to infer those non-integral shifts, and then super-resolve the images using various optimization frameworks. Depth estimation will be discussed in Section V. For the inference of non-integral shifts, Chan *et al.* [76] presented a mathematical model to simulate the super-resolution of light fields. By explicitly introducing Lambertian reflectance priors in the image formation model, Bishop *et al.* [77] formulated the reconstruction of the light field in a Bayesian framework, allowing more information to be recovered in the super-resolved light field. Mitra and Veeraraghavan [78] proposed a patch-based approach using a Gaussian

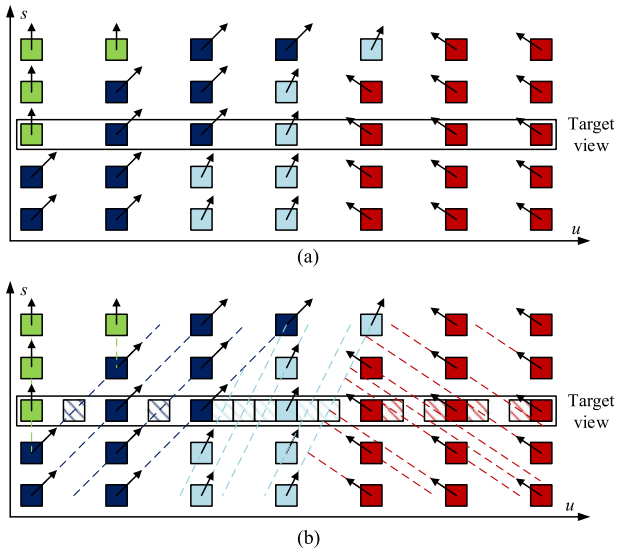


Fig. 9. Principle of light field super-resolution in the spatial dimensions. (a) Each pixel in an EPI is assigned a disparity label (indicated by the arrows) with subpixel accuracy. (b) The intensity values in other views are propagated to the target view. Due to the non-integral shifts between two corresponding pixels (the slope of the dashed lines in the figure), some pixels propagate to positions between the neighboring pixels (the pixels in the hatched-boxes), resulting in a super-resolved view.

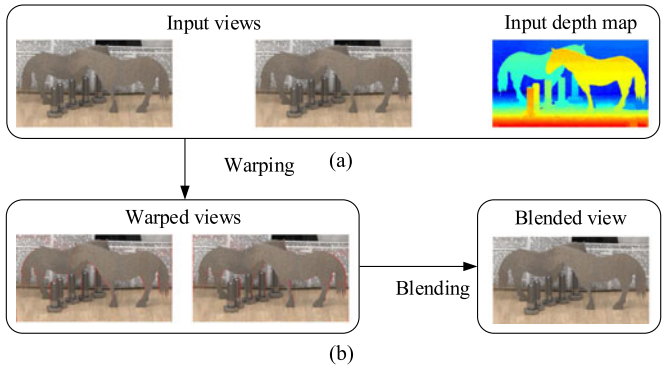


Fig. 10. Typical pipeline of depth image-based view synthesis. (a) Using the input depth map to warp the input images to novel views. In the novel, warped views, the red pixels are blank areas, also called "holes", caused by the warping operation. (b) Blending the warped views into a single image. The dataset in this figure is the *Horses* scene from the HCI database [74].

mixture model (GMM). They used a fast subspace projection technique to generate a disparity map and then modeled the light field patch as Gaussian random variables conditioned on its disparity value. Wanner *et al.* [66] used a structure tensor method to compute a continuous depth map on the EPI, then applied a variational optimization framework to super-resolve the images based on the depth map.

In contrast with the approaches above, which use the inherent information acquired by the plenoptic cameras, Boominathan *et al.* [67] proposed a hybrid light field imaging system that uses a single high resolution image captured by a DSLR camera to super-resolve the low resolution images in the light field. They extracted patches from the high resolution image with the lowest matching cost to those in the low resolution light field, and then blended the high resolution patches using a weighted average. Based on this approach, Wang *et al.* [68] applied an additional depth-based view synthesis method to warp the high resolution image to the position of each low resolution image, and built an iterative refinement pipeline to improve the warped images. Compared with the early work by Boominathan *et al.* [67], higher frequency information can be reconstructed in the low resolution images.

Taking advantage of convolutional neural networks (CNNs), Yoon *et al.* [72] presented a CNN-based approach to perform spatial super-resolution without depth estimation. Farrugia *et al.* [79] trained linear projections between subspaces of reduced dimension in which patch-volumes extracted from the light field reside.

B. Angular Super-Resolution

Many studies have focused on angular super-resolution using a small set of views with high spatial resolution. These views can be modeled as a reconstruction of the plenoptic function using limited samples. The existing angular super-resolution approaches can be divided into two categories: those that use depth estimation, and those that do not.

1) Depth Image-Based View Synthesis: Depth image-based view synthesis approaches typically first estimate the depth information and then warp the existing images to the novel view based on the estimated depth. The warped views are blended in a specific way. Fig. 10 illustrates this process.

Georgiev *et al.* [30] employed a segmentation-based optical flow method to compute the flow between an image and two neighboring views. Then, the novel view was generated by weighting the three warped views. Pearson *et al.* [80] introduced a depth layer-based method for synthesizing an arbitrary novel view. They assigned each pixel to a particular layer and then warped the pixels using a probabilistic interpolation approach. Wanner *et al.* [66], [81] formulated the view synthesis problem as an energy minimization one. Chaurasia *et al.* [82] presented a novel warping scheme using a superpixel technique, and then blended the warped views with weights specified by camera orientation and the reliability of depth information. Pujades and Devernay [69] proposed a view synthesis approach by optimizing a novel cost function with a Bayesian formulation that is robust to errors in the estimated depth. Zhang *et al.* [70] presented a phase-based approach using a micro-baseline stereo pair. They also introduced a disparity-assisted phase-based synthesis strategy to integrate the disparity (depth) information into the phase term when warping the input image to the novel view. Zhang *et al.* [83] proposed a layered patch-based synthesis method for multiple applications.

Typical depth-based view synthesis approaches rely heavily on the estimated depth, which is sensitive to textureless and occluded regions. In addition, they often focus on the quality of depth estimation rather than the quality of the synthetic views. In recent years, some studies based on CNNs aimed at maximizing the quality of the synthetic views have been presented. Flynn *et al.* [84] proposed a deep learning method to synthesize novel views using a sequence of images with wide baselines. Kalantari *et al.* [85] used two sequential CNNs to simultaneously estimate

depth and color by minimizing the error between the synthetic views and ground truth images.

2) Light Field Reconstruction without Depth: Depth image-based view synthesis approaches involve depth estimation, which tends to fail in occluded regions, as well as in glossy or specular ones. An alternative approach is based on sampling and consecutive reconstruction of the plenoptic function. For a light field that meets the requirement of minimum sampling described in Section II-C, novel views can be produced using light field rendering techniques [3].

For sparsely sampled light fields, direct interpolation will cause ghosting effects in the rendered views. To mitigate this effect, some studies have investigated light field reconstruction in the Fourier domain. Levin and Durand [86] proposed a linear view synthesis approach using a dimensionality gap prior to synthesize novel views from a set of images sampled with a circular pattern. Shi *et al.* [87] considered light field reconstruction as an sparsity optimization in the continuous Fourier domain. They sampled a small number of 1D viewpoint trajectories formed by a box and 2 diagonals to recover the full light field. Didyk *et al.* [88] used the phase information from a complex steerable pyramid decomposition to synthesize novel views with a small parallax. However, those reconstruction approaches always require the light field to be sampled in specific patterns, which imposes a restriction for practical usage. Instead, Vagharshakyan *et al.* [89], [90] considered angular super-resolution as an inpainting problem on the EPI, and used an adapted discrete shearlet transform to super-resolve the angular resolution of a light field.

Also, some learning-based methods have been presented for depth-free angular super-resolution. Yoon *et al.* [72] proposed a CNN-based approach that used two neighboring views in the vertical (horizontal) angular dimension to generate the middle view. However, they underused the potential of the entire angular information, and could only achieve a fixed super-resolution rate. Instead, Wu *et al.* [71] took advantage of the clear texture structure of the EPI and modeled the angular super-resolution as a CNN-based angular restoration of the EPI. Before feeding the EPI to the network, the spatial low-frequency component was extracted to avoid aliasing effects in the angular dimension of the EPI. Fig. 11 shows an angularly super-resolved result on a microscope light field [7]. In this case, a light field of 3×3 views was input to the network, resulting in a 7×7 views super-resolved light field.

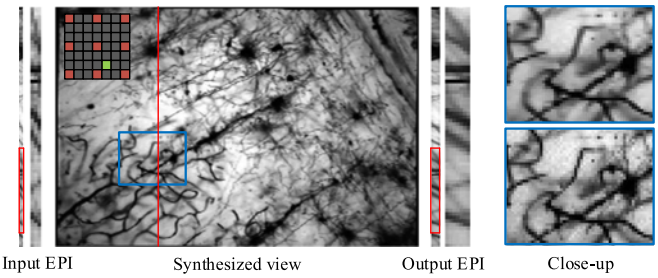


Fig. 11. Angular super-resolution on the Stanford microscope light field data *Neurons* $20\times$ [7] using 3×3 input views (output is 7×7 views). The angular super-resolution approach was proposed by Wu *et al.* [71], and models the reconstruction as a CNN-based angular restoration of the EPI (see the input EPI and the output EPI in the figure). The views in the blue boxes show the close-up version of a synthesized view (top) and corresponding ground truth view (bottom).

C. Temporal Super-Resolution

Due to the high dimensional data in a light field, current commercial light field cameras usually fail to capture light field video with a satisfactory frame rate. For example, Lytro Illum [4] captures a light field containing 7728×5368 pixels in a single shot, and can only achieve 3 fps in the continuous shooting mode. Most researchers have focused on video frame interpolation [91], [92]. But the problem becomes much more challenging when it comes to a light field video for the following reasons: First, the frame rate of a light field video can be extremely low; second, 4D light fields rather than 2D images need to be considered.

To super-resolve the temporal dimension of a light field, a hybrid imaging system composed of a plenoptic camera (Lytro Illum [4]) and a DSLR camera was developed by Wang *et al.* [73]. In this system, the plenoptic camera captured a light field sequence at 3 fps and the DSLR camera captured the corresponding 2D video at 30 fps. Then, a learning-based approach was designed to propagate the sequences containing a light field to the sequences containing only 2D images, and finally output a light field video at 30 fps. Specifically, a spatio-temporal CNN and an appearance CNN were designed, where the former was trained to warp the input images from the 2D video and the light field images to the target angular view, and the latter was trained to combine the warped images into the final image.

D. Spectral Reconstruction

Rather than directly capturing a hyperspectral light field, which would sacrifice the resolution of other dimensions, practical approaches split the task into 4D light field imaging and hyperspectral imaging [93]–[95], and then reconstruct the 5D light field in specific ways.

Wu *et al.* [93] presented a hyperspectral light field microscope that contains a 5×5 camera array, where each camera was mounted with a color filter (25 filters in total) to capture the scene at a certain spectral band. A 4D deconvolution algorithm is applied to reconstruct the hyperspectral light field from the spatial-spectral coupled sampling. Instead of using a camera array and taking advantage of a hybrid imaging approach, Xiong *et al.* [94] developed a hyperspectral light field imaging system using a plenoptic camera and a coded aperture snapshot spectral imager (CASSI). In this system, the incident light was divided by a beam splitter and then captured by the plenoptic camera and CASSI, where the plenoptic camera recorded a 4D light field and RGB spectral information, and the CASSI encoded a 2D image with 27 spectral bands. To reconstruct the final hyperspectral light field, self-learned dictionaries were trained to exploit the large correlations across the angular and spectral dimensions.

TABLE II
COMPARISON OF LIGHT FIELD SUPER-RESOLUTION APPROACHES

Approach	Main feature	Domain	Evaluation (PSNR)					
			Synthetic data			Real-world data		
			Buddha	*Mona*	*Still life*	*Maria*	*Couple*	*Tarot*
Wanner *et al.* [66]	Variational optimization	Spatial (2×)	34.50	–	24.93	35.18	25.54	26.66
PaSR [67]	Hybrid input, patch-based	Spatial (8×)	–	–	25.11	34.30	27.38	30.77
iPADS [68]	Hybrid input, patch- and depth-based	Spatial (8×)	–	–	26.45	36.10	28.65	31.63
Wanner *et al.* [66]	Variational optimization	Angular (using 8 views)	42.84	–	30.13	40.06	26.55	28.71
Pujades *et al.* [69]	Robust to depth	Angular (using 8 views)	42.37	–	30.45	40.10	28.50	28.88
DAPS [70]	Phase-based, micro-baseline	Angular (1 × 2 to 9 × 9)	44.71	–	34.81	40.81	32.07	–
Wu *et al.* [71]	Learning-based	Angular (4×)	43.20	44.37	–	–	–	–
LFCNN [72]	Learning-based	Spatio-angular (2×)	36.86	37.56	–	–	–	–
Wang *et al.* [73]	Learning-based	Temporal (10×)	–	–	–	–	–	–

The evaluation provides numerical results of super-resolved light fields using the listed implementations. For angular super-resolution approaches by Wanner *et al.* [66] and Pujades *et al.* [69], 8 views were used to synthesize each novel view. The datasets *Buddha*, *Mona*, *Still life*, *Maria* and *Couple* are from the HCI light field dataset [74], and *Tarot* is from the (New) Stanford light field archive [24].

E. Discussion

Resolution trade-offs being an intrinsic problem of light field acquisition, researchers have devoted great attention to light field super-resolution. Table II lists numerical evaluation (PSNR) of some representative light field super-resolution approaches. From these numerical results and the discussed techniques, we can draw the following conclusions:

1) Light field super-resolution can greatly benefit from hybrid imaging systems; these have several merits, such as much smaller data size and compressibility. Typical approaches, such as PaSR [67] and iPADS [68], have an extremely large super-resolution rate.
2) Approaches based on learning techniques (e.g., CNNs, or sparse coding) can achieve better performance, especially when taking advantage of properties of light fields structure, such as the EPI structure. Among numerous super-resolution techniques, the learning-based approaches, e.g., Wu *et al.* [71] and Kalantari *et al.* [85], currently constitute the state of the art.
3) Super-resolution of non-Lambertian regions can be a great challenge, especially for a microscopy light field, where the depth information is difficult to obtain.

V. LIGHT FIELD DEPTH ESTIMATION

In this section, we present techniques devoted to depth map estimation from light field data. The four-dimensional light field representation essentially contains multiple views of the scene, making depth map estimation possible. In comparison to stereo vision-based methods, light field-based methods do not require camera calibration, making it convenient for data acquisition in real-world scenes. Moreover, the peculiar structure of the light field expands the disparity space to a continuous space [66], making depth estimation more robust and precise. However, processing this high dimensional data requires large computing times and resources, which is still a barrier for real-time depth estimation from light fields.

Typical light field depth estimation approaches first estimate an initial depth map in specific ways, then global optimization frameworks or local smoothing approaches are used to refine the depth map. Fig. 12 demonstrates this process. In this section, we will investigate light field depth estimation following this scheme.

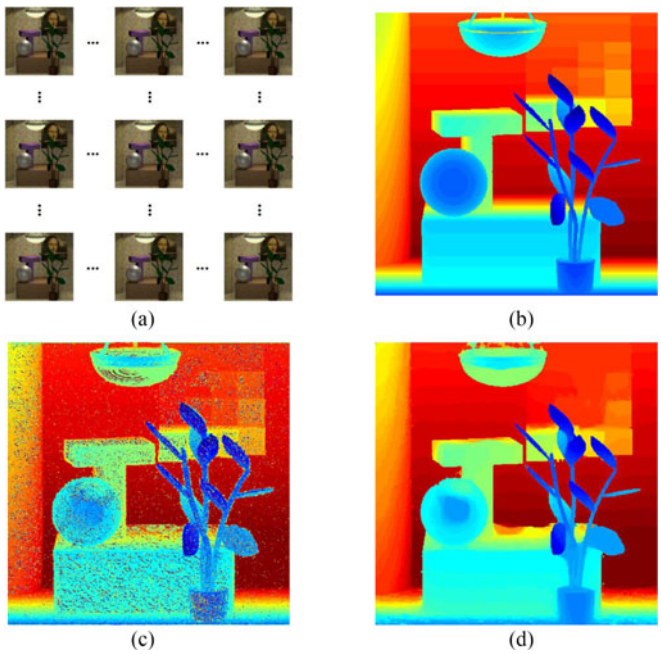


Fig. 12. Typical depth estimation pipeline using light field data. (a) Input light field. (b) The ground truth used for comparison with the final result. (c) Initial estimated depth map (please refer to Section V-A), which contains outliers due to noise, occlusions and inherent matching uncertainty caused by textureless regions. (d) Final depth map after depth refinement (please refer to Section V-B). The depth map is refined by means of a MRF framework (see the discretization effect in the background). The scene in this figure is *monasRoom* from the HCI database [74]. (a) Light field (b) Ground Truth (c) Initial Depth Map (d) Final Depth Map.

A. Initial Depth Estimation

The baseline between adjacent views in a light field is narrow, which makes it difficult to recover disparity from two views using traditional stereo matching methods. Therefore, instead

of using stereo matching methods, constraints and cues which take advantage of all the views together are used to estimate the depth map from a light field image. Existing initial depth estimation approaches can be divided into three categories: sub-aperture image matching-based methods, EPI-based methods and learning-based methods.

1) Sub-Aperture Image Matching-based Methods: As described in Section II-B, a sub-aperture image is recorded by light rays arriving at one point in the st plane. Each pair of sub-aperture images has a very narrow baseline, making the disparity range of sub-aperture images also quite narrow. For example, the disparity range between adjacent sub-aperture views of the Lytro camera is smaller than 1 pixel [96]. With such narrow baselines, the sub-pixel shift in the spatial domain usually involves interpolation with blurriness, which leads to poor performances on correspondence-matching approaches. Moreover, the matching costs of stereo correspondence are highly ambiguous.

Therefore, instead of using stereo matching, different constraints that leverage the structure of light fields are used to estimate the initial depth map. Different from stereo matching based on two views, all views are contained in the constraints. Yu *et al.* [96] explored the geometric structure of 3D lines in ray space to improve the light field triangulation and stereo matching. Since the method uses stereo matching to first estimate each pixel's disparity, it does not perform well on the datasets with small disparity range. To solve the problem, Heber *et al.* [97] proposed a novel principal component analysis (PCA) matching term for multi-view stereo reconstruction. They assumed that if one were to warp the images of various views to a common warping center and then consider each warped image as one row in a matrix, the matrix should be low rank. In order to match sub-aperture images with an extremely narrow baseline, Jeon *et al.* [98] presented an algorithm based on the cost volume. The cost volume is computed to evaluate the matching cost of different disparity labels, which is defined using the similarity measurement between the sub-aperture images and the center view sub-aperture image shifted at different sub-pixel locations. To generate a better result on occlusions, Wang *et al.* [99] developed a novel occlusion scheme for their occlusion-aware depth estimation algorithm. By ensuring photo-consistency in only the occluded view region, their method enabled identification of occluded edges, a technique that could be useful in other applications. Based on this occlusion model, Zhu *et al.* [100] extended the approach above to the context of multi-occluder occlusion. By using the concept of the surface camera (SCam), which models angular radiance distribution with respect to a 3D point, Chen *et al.* [101] introduced a bilateral consistency metric on SCams for light field depth map estimation.

Also, various other cues can be exploited in depth estimation, the defocus cue being the most prominent one. Ng's work [28] noted that a light field image can be refocused on any area by shearing the light field using a combination of digital refocusing and Fourier refocusing. After refocusing at a new depth, the light rays from any in-focus scene point will be recorded in the same spatial location—a concept also called angular coherence (shown in Fig. 13). We can refocus the light field at different depth candidates. The absolute difference between the central view and the other angular views reflects the depth probability.

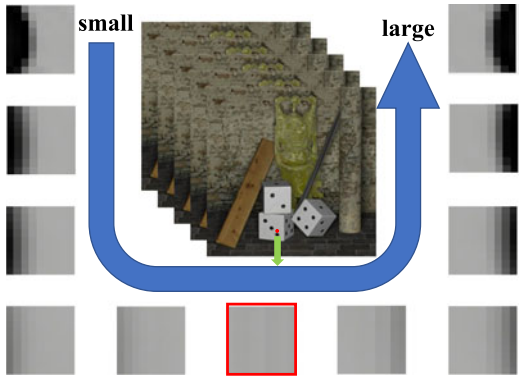


Fig. 13. We display a patch, consisting of all views recorded for a point, as would appear under different depth candidates. The red point marked in the light field view is the one selected. In counter-clockwise order, the depth values of the patches range from smaller to larger depths. Among these candidates, since the pixels in the patch marked with a red frame have the largest similarity among them, it is that depth the one that has the largest probability of being the actual depth of that point.

Based on Ng's work [28], Tao *et al.* [102] combined the defocus cue and the correspondence cue to estimate the depth of scenes. In addition, they extended their analysis, using the additional cue of shading, to refine fine details in the shape [103]. For light fields including a scattering medium, such as underwater scenes or scenes with fog, Tian *et al.* [104] combined the correspondence and defocus cues with a novel transmittance-based depth cue to remove the influence of backscatter and attenuation. Williem *et al.* [105] proposed two novel cues, which are an angular entropy metric and the adaptive defocus response, to improve the robustness of depth estimation to occlusion and noise. Lin *et al.* [106] presented two features for depth estimation: one feature is the property that non-occluding pixels exhibit symmetry along the focal depth dimension centered at the in-focus slice; the other is a data consistency measure based on the analysis of a synthesized view focused at a hypothesized depth. Lee and Park [107] proposed a depth estimation approach by accumulating binary maps, which are obtained by separating the foreground and background with respect to a series of focused planes.

For the particularly challenging case of depth estimation for glossy or specular surfaces, Tao *et al.* [108] presented a novel theory of the relationship between light-field data and reflectance in the dichromatic model. Wang *et al.* [109] derived a spatially-varying (SV) BRDF-invariant theory for recovering 3D shapes and reflectance from light field cameras to estimate the depth of non-Lambertian planes. They proposed an equation relating depths and normals that could be derived to estimate the depth map.

2) EPI-based methods: Instead of performing stereo matching on sub-aperture images, novel approaches considering the properties of the light field structure have been developed. Before the rise of light field cameras, epipolar geometry was used to achieve scene reconstruction from images taken from different views. After the introduction of the light field camera, EPIs

could be generated directly from light field images; in them, the slopes of the lines are indicative of the depths of the different objects (as we mentioned in Section II-B). Most EPI-based depth map estimation approaches rely on measuring the slopes in EPIs using different optimization techniques.

In 1987, Bolles *et al.* [12] presented a technique for building a 3D description of a static scene from a dense sequence of images. The technique utilized knowledge of the camera motion to form and analyze slices of the solids generated by these images. Criminisi *et al.* [115] addressed the problems of achieving a dense scene reconstruction from a sequence of images, as well as analyzing and removing specular highlights. The key was to directly exploit the high degree of regularity found in the EPI volume. Basha *et al.* [116] proposed a method for estimating the 3D structure and the dense 3D motion (scene flow) of a dynamic non-rigid 3D scene, using a camera array. Their core idea was to use a dense multi-camera array to construct a novel, dense 3D volumetric representation of the 3D space in which each voxel holds both an estimated intensity value and a confidence measure of that value.

Kim *et al.* [25] introduced a method to perform scene reconstruction of complex, detailed environments from high angular resolution EPIs. Their method computed the 3D depth information for all the visible scene points. They applied a confidence measure in EPI space to compute the reliability of the estimated depth. Wanner *et al.* [66] utilized the structure tensor in EPI space to estimate the local direction of a line. Based on their work, Li *et al.* [110] made use of the structure information to build a reliability map for EPIs. Krolla *et al.* [111] extended depth estimation using the structure tensor to spherical light fields. Diebold *et al.* [113] modified the structure tensor approach for heterogeneous light fields in which the EPIs have non-constant intensity in each line. Ziegler *et al.* [117] extended depth estimation in EPI space to the 4D epipolar volume. To handle highly noisy and strong occlusions, Zhang *et al.* [118] integrated a spinning parallelogram operator in the 2D EPI, where the operator measures the slopes of an EPI by maximizing distribution distances between two parts of the parallelogram window.

3) Learning-based methods: More recently, researchers have presented techniques for depth map estimation based on CNNs. Heber *et al.* [114] utilized CNNs to predict depth information for light field data learning an end-to-end mapping between the 4D light field and a representation of the corresponding 4D depth field in terms of 2D hyperplane orientations. With the development of sparse coding, Johannsen *et al.* [119] proposed a novel method for depth estimation in light fields that employed a specifically designed sparse decomposition to leverage the depth-orientation relationship in EPIs.

B. Depth Refinement

Due to noise, occlusions or inherent matching uncertainty caused by textureless regions, the initial depth map typically contains outliers, see Fig. 12(c). Therefore, a depth map refinement is crucial to smooth out the outliers and produce the final depth map. Typical methods are either within a Markov random field (MRF) framework or a variational framework.

MRF are widely used for problems that involve feature point matching [96]–[99], [102]. The framework includes a unary term and a pairwise term. The unary term is first constructed using initial depth estimation approaches described above, then discretized to different depth values. Each unary term value reflects the probability of the corresponding depth. However, in textureless areas, the unary term is easily impacted by noise, making it inaccurate at estimating the depth of such areas. To address this, a pairwise term was added to refine noise points with their neighbors. The pairwise term is designed using the properties of the depth map, such as smoothness or color consistency. An energy function combines the unary term and the pairwise term, and the depth map is estimated by minimizing it. Fig. 12(d) shows an example depth map refined with a MRF framework.

Although MRF-based refinement performs well, it results in *discrete* disparity values (see the background region in Fig. 12(d)). Consequently, this depth map refinement approach does not perform well for 3D reconstruction. To address the problem, Wanner and Goldluecke [66], [112] presented a continuous framework and described novel variational methods for disparity reconstruction. Their framework employed a total variation (TV) smoothing scheme. Based on their work, Heber *et al.* [120] proposed a model that combines the main idea of active wavefront sampling (AWS) with light fields. By tracking an image scene point over a sequence of sub-aperture images, they observed a virtual rotation of the scene point on the image plane, that was inversely related to the scene depth. To construct their model, they utilized an extension of the total generalized variation (TGV) approach.

C. Discussion

In this section, we have introduced depth estimation approaches based on light field data. Table III lists numerical evaluation values (MSE) of some representative approaches. These methods help to generate accurate depth maps based on light fields, but challenges remain. First, most methods are based on depth estimation of Lambertain surfaces; while for non-Lambertian ones, in which a change in viewing direction can imply a drastic change in appearance, traditional constraints and cues do not work so well. Second, the resulting depth map can be quantized, leading to errors on it. Third, due to the constraint on the angular resolution of the light field image, there exists a restriction for the depth range.

Compared with stereo pairs, a light field provides a more continuous disparity space due to a denser viewpoint sampling, i.e., higher angular resolution. Intuitively, a higher angular resolution of the light field benefits depth estimation. However, from the point of view of hardware design, higher angular resolution means denser sensor arrangement (in the case of multi-sensor capture), longer capture time (in the case of time-sequential capture), or lower spatial resolution (in the case of spatial multiplexing). Besides, dense viewpoint sampling also requires accurate calibration tools [121] or mechanical controllers. In practice,

TABLE III.
COMPARISON OF LIGHT FIELD DEPTH MAP ESTIMATION APPROACHES

Category	Approach	Main feature	Evaluation (MSE)					
			Buddha	*MonasRoom*	*Still life*	*Horses*	*Medieval*	*Papillon*
Matching-based	Wang *et al.* [99]	Occlusion-aware	2.06	0.87	7.24	1.76	3.62	0.43
	Jeon *et al.* [98]	Sub-pixel accuracy	2.75	0.43	1.52	4.91	5.34	1.9
	Yu *et al.* [96]	Line-assisted graph cut	5.03	18.8	12.88	18.8	28.45	17.29
	Heber *et al.* [97]	PCA matching term	–	–	–	–	–	–
	Tao *et al.* [102]	Defocus cues	–	–	–	–	–	–
EPI-based	Kim *et al.* [25]	Large scene reconstruction	–	–	–	–	–	–
	Li *et al.* [110]	Sparse linear optimization	0.64	0.64	3.25	0.95	2.1	2.28
	Krolla *et al.* [111]	Spherical light field	–	–	–	–	–	–
	Wanner *et al.* [112]	Total variation	–	–	–	–	–	–
	Diebode *et al.* [113]	Modified structure tensor	–	–	–	–	–	–
Learning-based	Johannsen *et al.* [66]	Sparse coding	0.57	0.65	2.95	3.26	0.84	1.85
	Heber *et al.* [114]	CNN-based	–	–	–	–	–	–

These approaches are classified into different categories according to their initial depth estimation approaches. The evaluation provides numerical results of MSEs using the listed implementations. The datasets *Buddha*, *MonasRoom*, *Still life*, *Horses*, *Medieval* and *Papillon* are from the HCI light field dataset [74].

researchers have found that the quality of the estimated depth can be greatly increased when using an angularly super-resolved light field [66]. The effectiveness can be more significant when the light field is super-resolved using a depth-free framework, such as learning-based methods [71], [72]. Conversely, a depth map of better quality, such as with better occlusion handling, produces better super-resolution results. One may model depth estimation and light field super-resolution in a joint optimization framework to obtain both a high quality depth map and a super-resolved light field.

VI. LIGHT FIELD ASSESSMENT AND COMPRESSION

Light field quality assessment and light field compression are two critical aspects for the practical usage of light fields. Light field assessment is helpful to obtain a better understanding of the performance of light field acquisition, generation, and processing techniques. With regard to compression, a light field involves a large amount of data, but also records a scene with a set of images from different viewpoints, thus exhibiting data redundancy in both the spatial and angular dimensions [3]; examples of this redundancy are the smooth regions in each sub-aperture image (Fig. 2(a)) and light field subview (Fig. 2(b)). Therefore, the compression approaches for light fields offer much opportunity and are essential for light field storage, transmission and display. In this section, we introduce quality assessment approaches and compression schemes for light fields.

A. Quality Assessment

Light field quality assessment usually involves assessment of spatio-angular resolution, spatial quality, and angular consistency. Existing image resolution assessment methods adopt the ISO 12233 standard [122], which uses a slanted edge-based method to calculate the spatial frequency response. Fu *et al.* [123] showed that a light field camera exhibits more stabilized visual resolution in terms of depth of field (DOF) compared with a conventional camera due to the refocusing capability. For spatial quality assessment, peak signal to noise ratio (PSNR) and structure similarity index (SSIM) [124] are always applied. For instance, Shidanshidi *et al.* [125] applied PSNR to evaluate synthesized novel views using different ray interpolation approaches [3] with the corresponding simulated views. Angular consistency measures the visual coherence between views. Such criterion is always critical to the evaluation of light field processing approaches, such as angular super-resolution and light field editing. Current angular consistency assessment usually relies on subjective evaluations on an EPI or a pseudo sequence. Adhikarla *et al.* [126] presented an interactive light field viewing setup for the subjective evaluation of angular consistency. Besides, they extended SSIM [124] to a 3D context for light field-specific angular assessment, and evaluated the performance of existing quality assessment metrics.

B. Light Field Compression

In this subsection, we investigate lossy and lossless compression schemes for light fields; the lossy compression is further divided on the following taxonomy: progressive or transform coding, predictive coding and pseudo-sequence-based coding. Note that some techniques employ hybrid coding schemes; we classify these techniques according to their main aspect.

1) Lossy Compression: **Transform coding approaches** typically rely on a transform, such as the discrete cosine transform (DCT) or the discrete wavelet transform (DWT), to compress a light field. Classical coding schemes, such as JPEG (using DCT) and JPEG 2000 (using DWT), are applied to compress the light field raw image. However, these two schemes are not specifically designed for light field compression, and will thus not achieve optimal compression results. Aggoun [127] proposed a 3D-DCT to exploit the spatial redundancy within the light field subviews. Specifically, they first arranged a selected subview and its neighboring ones into a 3D brick, then the 3D-DCT was applied to produce the decorrelated subview group. Rather than 3D-DCT, Aggoun [128] further applied a 3D-DWT to a brick of sub-aperture images, and Magnor *et al.* [129] presented

4D-DWT to directly compress the 4D light field without arranging subviews or sub-aperture images.

Those approaches may produce misalignment between views due to large displacements between neighboring views, and the resulting high frequencies are expensive to code. To handle this problem, disparity compensation was applied to warp each view onto its view-dependent texture map. Xu *et al.* [130] first decomposed a light field into sub-bands using wavelet packet transform; the wavelet packet bases were divided into two groups: predictable bases and unpredictable bases. Finally, a disparity map, which was applied for sub-band prediction, was estimated using an optical flow-based approach. Chang *et al.* [131] incorporated disparity compensation into wavelet decomposition based on lifting. They first applied disparity-compensated wavelet decomposition to exploit similarity between different views in the light field, then the resulting coefficient images were encoded using shape-adaptive discrete wavelet transform.

Predictive coding approaches first select a set of images from the light field array to be coded as *Intra*, also known as I-images. Then the I-images serve as reference for coding the remaining light field images, also known as P-images. An early work proposed by Magnor and Girod [132] first separated each sub-aperture image into square blocks, and predicted the blocks in the P-images using the blocked disparity and blocks in the I-images. Similarly, Conti *et al.* [133] used blocked disparity for the coding of 3D light fields and 3D light field video. Based on Magnor and Girod's work, Kundu [134] used 2D homography to match P-images with I-images, and predicted P-images using homography-based warping and interpolation. Jiang *et al.* [135] further combine homography and low rank approximation, which are jointly optimized. By considering the similarity between adjacent views, Liu *et al.* [136] proposed a layered prediction scheme that predicts the P-images layer by layer, i.e., views in a lower layer were applied to predict views in a higher layer. For large compression rates, Li *et al.* [137] presented a coding scheme that uses a sparse set of subviews and associated disparities to compress the full light field. The coding scheme first applied the sparse set of subviews and disparities to reconstruct an intermediate light field, then utilized it to predict the original light field using a modified high efficiency video coding (HEVC) encoder. To improve coding efficiency, Jin *et al.* [138] employed a reshaping algorithm before the compression.

Pseudo sequence coding. A 4D light field can be considered as multiple frames with slightly different viewpoints. Based on this assumption, pseudo sequence-based coding approaches first rearrange light field elements (usually sub-aperture images) as a pseudo sequence, then use a video encoder to compress the sequence. Typical rearrange paths are *zig-zag* and *spiral* [139], [140], where the *zig-zag* scan path rearranges the views from left to right and top-down (as shown in Fig. 14(a)) and the *spiral* scan path begins from the center view and progresses from the inner to the outer ones spirally (as shown in Fig. 14(b)). For the compression, Dai *et al.* [139] and Vieira *et al.* [140] applied standard video encoders, such as H.264/AVC and HEVC, to compress the pseudo sequence. In their experiment results, the *spiral* scan path outperforms the *zig-zag* pattern due to a smoother rearrangement. Rather than using the scan paths

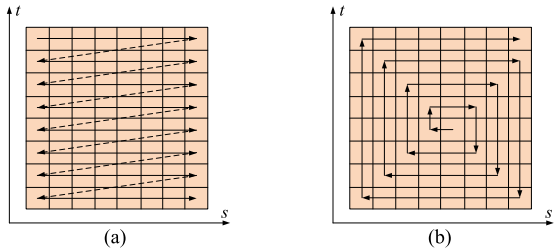


Fig. 14. Typical rearrangement paths for pseudo sequence-based coding approaches: (a) *zig-zag*; and (b) *spiral*.

described above, Li *et al.* [141] proposed a 2D hierarchical coding structure to reduce the reference buffer. They divided the views into four quadrants, and encoded one quadrant after another.

2) Lossless Compression: Perra [142] presented a lossless compression scheme on raw plenoptic data. They first divided the raw image into four color components and then encoded them separately using an adaptive prediction tool. Instead of encoding raw plenoptic data, Helin *et al.* [143], [144] implemented their compression scheme on rectified light field images, and thus, it was not limited to plenoptic data. By exploiting the high redundancy existing between the sub-aperture images, their lossless compression scheme first constructed consistent segmentations for the central view, and then propagated them to other views using a quantized depth map. An optimal sparse predictor was then applied to predict each region of the segmentations using the corresponding regions from the already coded views.

3) Discussion: A light field contains high dimensional data, posing a great challenge for practical usage, storage, transmission and display. Fortunately, light fields exhibit data redundancy in both the spatial and angular dimensions [3], which can be effectively removed using compression techniques. Some events, such as JPEG Pleno activity [145] and ICME 2016 Grand Challenge on Light-Field Image Compression [146], were held to look for effective compression schemes. Current stat-of-the-art techniques can achieve more than 80% bitrate reduction [137] for lossy compression, and 60% bitrate reduction [143] for lossless compression. However, for practical usage of light fields, more progress is expected for light field compression schemes that aim at providing high compression quality with reasonable resource requirements in terms of data rates, computational complexity, and power consumption. Besides, in lossy compression schemes, the coding and decoding process for the residual will distort the structure of the input 4D light field to a certain extent. This structure distortion will definitely influence the subsequent light field processing, such as depth estimation or super-resolution. Thus, obtaining a decent bitrate reduction while preserving the light field structure is also a challenge for light field compression [147]. A more detailed evaluation of state-of-the-art light field compression approaches can be found in [148].

VII. Light Field Applications

When the light field was first introduced to computer graphics in 1986 [151], researchers proposed only one application:

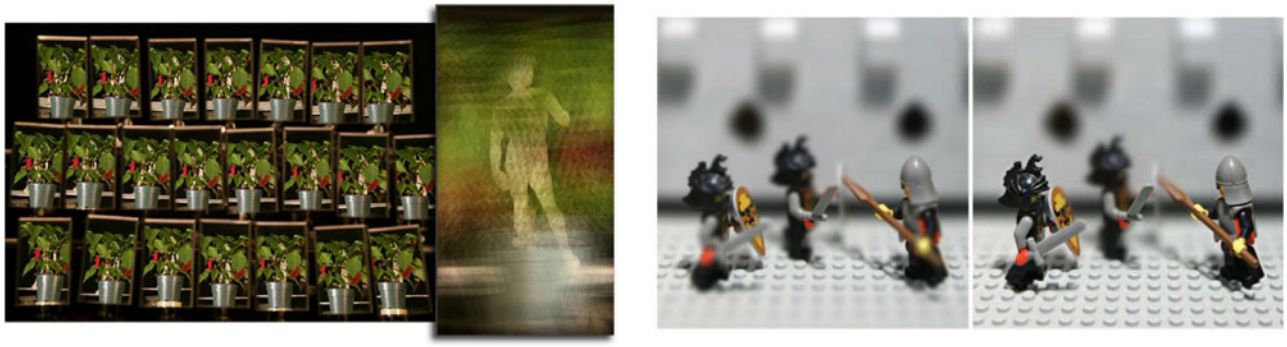


Fig. 15. Applications using light fields. (a) The light field acquired by capturing an array of planar mirrors (left) was used to generate a synthetic aperture image (right) [149]. (b) A light field can be applied for post-capture refocusing on a single plane (left) or over a controllable range of depths (right) [150].

novel view synthesis. However, as computing speed and computational resources increased, the light fields have grown in popularity in computer vision applications. In this section, we investigate various applications that benefit from light field data.

A. Synthetic Aperture Imaging

Isaksen *et al.* [36] first demonstrated the light field's ability to "see-through an occluder" by synthesizing a sufficiently large aperture. Because the object of interest and the occluder are at different depth layers and parallaxes, each view is able to capture a small portion of the object [152]. By propagating the portion of interest in each view, one can synthesize a virtual aperture image at a specific depth layer.

Levoy *et al.* [149] used synthetic apertures to reconstruct partially obscured objects (e.g., seen through a plant). A single camera was used to capture images from different viewpoints reflected by an array of planar mirrors. With an appropriate focal plane, a virtual view with a wide aperture and a shallow DOF was synthesized, where the object was in focus and occluders were blurred out (see Fig. 15(a)). Vaish *et al.* [153] studied four cost functions, including color medians, entropy, focus and stereo, for reconstructing occluded surfaces using a synthetic aperture. The cost functions, however, may fail under severe occlusions. Yang *et al.* [154] divided the scene into multiple visibility layers, and then propagated the visibility information among those layers using an optimization framework. Their approach was able to produce a synthetic aperture image with a wider DOF compared with previous works. Dansereau *et al.* [150] presented a volumetric focus approach that improves focus quality while maintaining sharp focus over a selected depth range using light fields (see Fig. 15(b)).

In addition, some studies for tracking objects through occlusions based on synthetic aperture techniques have been proposed [155], [156]. Occlusions are one of the obstacles for object tracking. Conventional single-camera tracking algorithms typically predict the motion parameters using filtering techniques such as Kalman filtering or particle filtering. However, the tracking tends to fail when the motion trajectory is unpredictable. Rather than tracking an object in a passive way, a camera array system (e.g., Joshi *et al.* [23] used 8 VGA video cameras) is able to provide clearer views of objects of interest using synthetic aperture techniques.

B. Segmentation and Matting

Typically, segmentation can be considered as a labeling problem. Unlike conventional segmentation approaches, such as co-segmentation and multiview segmentation, views in a light field are much more correlated (i.e., EPI volume), and thus can help in ensuring labeling consistency. However, due to the large number of views in a light field (usually a dozen [157] to a hundred views [158], [159]), segmentation and matting using light field data take abundant computing resources and running time during the energy minimization, which restricts their further promotion.

Campbell *et al.* presented multiple-view segmentation approaches using graph-cuts in voxel space [161] and image space [162]. However, the MRF framework is always time-consuming. To handle this problem, Hog *et al.* [163] proposed a graph structure to reduce the graph size of a light field. By taking advantage of structure information in the EPI space, Berent and Dragotti [164] modeled the segmentation problem as an energy minimization scheme. Wanner *et al.* [157] presented a variational framework for multi-label segmentation on the ray space of 4D light fields. Yücer *et al.* [158], [159] proposed a segmentation approach that uses unstructured 3D light fields captured by a hand-held video camera (see Fig. 16(a)). They computed the gradients in the light field volume and used local gradient filters to estimate the likelihood of edges belonging to the foreground. The likelihood was then propagated to the remaining areas in the volume. Finally, a Bayesian framework was employed to aggregate the segmentation likelihood of every image into the 3D volume. Based on the estimated depth, Zhu *et al.* [165] propagated the superpixel segmentation in a sub-aperture image to the whole light field to form a light field superpixel segmentation. Also, light fields have been used to segment refractive surfaces or transparent objects into different superimposed layers. Johannsen *et al.* [166] used disparity and luminance information from a light field and formulated a linear generative model to segment the light field into individual layers, while Xu *et al.* [167] proposed a transparent object segmentation approach using the consistency and distortion properties of a light field.

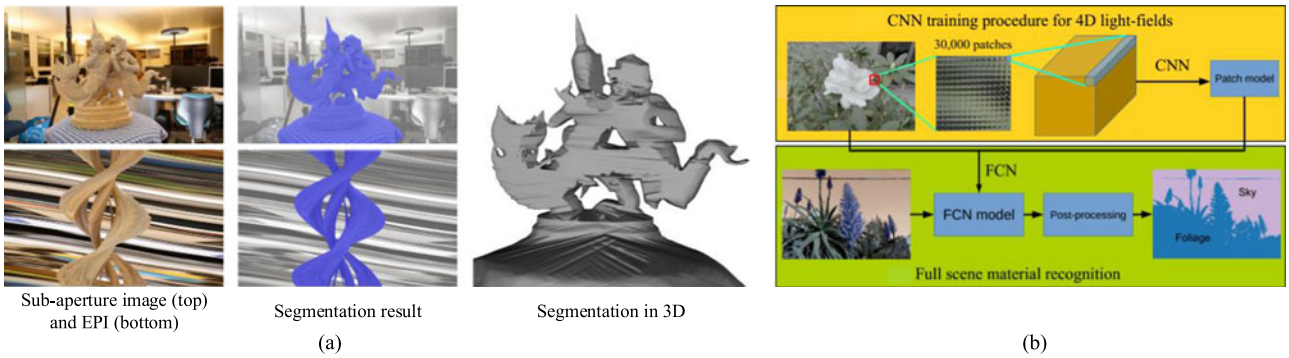


Fig. 16. Further examples of applications using light fields. (a) Object segmentation proposed by Yücer *et al.* [158], [159], where the likelihood of edges belonging to the foreground was estimated using the gradients in the 3D light field volume. The left column shows a sub-aperture image and an EPI of the light field, the middle column shows the segmentation results for them, and the right column shows the 3D segmentation result. (b) Material recognition proposed by Wang *et al.* [160], where a CNN was trained on a 4D light field dataset.

To solve the matting problem using a light field, Joshi *et al.* [168] proposed an efficient algorithm for natural video matting using a camera array. By taking advantage of light field data, a synthetic aperture image focused on the foreground was computed to reduce the variance in the foreground. The variance measurement was then used to construct a trimap for the alpha matte. Cho *et al.* [169] built their framework on the EPI to minimize the linear composite error, which is subject to correspondence constraints. Fiss *et al.* [170] used light field matting to separate foreground occluders (e.g., dust on a window or a wire mesh) from the background target. They represented the 4D light field as a foreground layer composited over a background light field.

C. Detection, Classification, and Recognition

Due to the inherent robustness to occlusions, light fields have an excellent performance on detection and recognition problems. Li *et al.* [171] demonstrated that light fields can benefit saliency detection. Their approach capitalized on the refocusing capability of light field imaging to measure likelihood scores of foreground and background. Compared with single-image saliency detection, their approach can more robustly handle challenging scenarios such as similar foreground and background.

Also, Pei *et al.* [172] used a camera array to synthesize a confocal view for multi-object detection. Ralph *et al.* [173], [174] introduced a novel face recognition approach using a light field system composed of 13 cameras. Raghavendra *et al.* [175] proposed a face detection approach from the best-focused image based on a wavelet energy analysis, and Raja *et al.* [176] used the best-focused image rendered by a light field camera for iris recognition.

Compared with object recognition, material recognition is a challenging problem because of the sheer diversity in the appearance of materials. Using the recently developed learning techniques, Wang *et al.* [160] proposed a CNN-based material recognition method that used a 4D light field dataset (see Fig. 16(b)). They also presented several CNN architectures for the 4D light field inputs as well as a light field dataset for material recognition. Compared with conventional single (2D) image material recognition approaches, their approach achieved approximately a $6-7\%$ boost.

D. Handling Refractive and Transparent Object

Light field-based detection and reconstruction of non-Lambertian materials [177], [178], such as refractive and transparent objects, has been studied, a problem that is extremely ill-posed when only a single image is considered. Maeno *et al.* [179] presented a transparent object recognition approach using a light field camera. They observed that a transparent object deforms background scenes due to refraction, causing ray distortions in the 4D light field. They named this observation the "light field distortion feature", and used it to detect transparent objects. Anglin *et al.* [180] proposed a shift-invariant refocusing transformation for volumetric estimation, which can be further applied for reconstruction of special materials, such as fluid and flame. Ding *et al.* [181] used the distortion of a known pattern caused by refraction to reconstruct a fluid's surface. They applied a camera array to track distorted feature points over time and across cameras and built spatio-temporal correspondence maps for the fluid surface reconstruction. Sulc *et al.* [182] used a 4D light field to separate a dichromatic reflection component from diffuse object colors. They model the observed light field as a sparse linear combination of a constant-color specular term and a small finite set of albedos. Lu *et al.* [183] considered light field sampling as sampling a 4D BRDF, and presented a CNN-based architecture for BRDF identification.

E. Visual Odometry, Localization and SfM

Dansereau *et al.* [184] presented a six-degree-of-freedom visual odometry approach that used a camera array to estimate the pointwise plenoptic flow. Their approach achieved a higher performance than conventional approaches (e.g., stereo feature-tracking) in terms of both robustness and tracking accuracy. Instead of performing stereo matching between images, Zeller *et al.* [185] presented a visual odometry approach that is able to find stereo correspondences in full sensor resolution, avoiding aliasing effects due to spatial undersampling. Schillebeeckx and

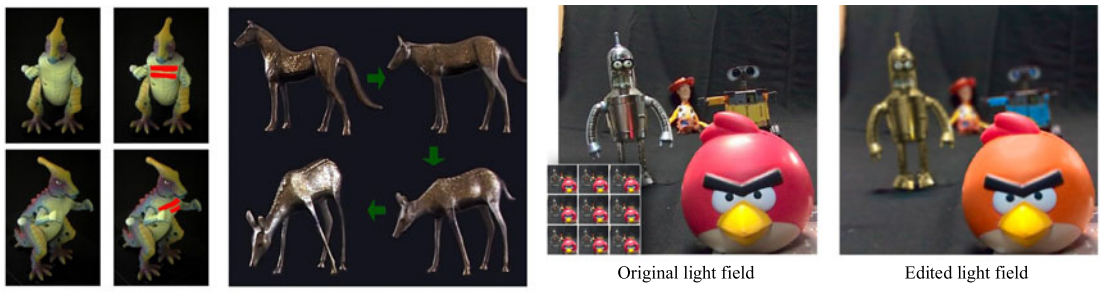


Fig. 17. Example edits on light fields. (a) Seitz and Kutulakos [193], [194] used a voxel-based representation of the light field, allowing drawing and scissoring. (b) Light field morphing between two input light fields [195]. (c) Sample light fields captured with the Lytro camera, edited using the different paradigms analyzed by Jarabo *et al.* [196].

Pless [186] used light fields to simplify the estimation of camera rotation. They applied a lenticular sheet with a color-coded backplane pattern to create a light field where the observed color depends on the direction of the light. Srinivasan *et al.* [187] developed an oriented 4D light field window for robust and accurate pixel comparisons, and applied it to scene flow recovery. Fahringer *et al.* [188] further reduced complexity necessary to acquire volumetric data and used the 3D flow for velocity field estimation.

F. Stitching, Deblurring and Stabilization

Guo *et al.* [189] proposed a multiple light fields stitching algorithm to produce a seamless navigation using a light rays transformation matrix, named "ray-space motion matrix". Similarly, Johannsen *et al.* [190] used ray-to-ray correspondence to produce a light field panorama. Srinivasan *et al.* [191] took advantage of the 3D capture ability of the light field camera, introducing a general model for 3D camera motion estimation and deblurring. The model produced higher quality light field deblurring results when compared with conventional 2D image deblurring, which typically assumes the blur kernel to be a 2D function. Smith *et al.* [192] described a video stabilization approach using an integrated camera array, ProFUSION-25C [21]. They computed a sequence of relative poses between the virtual camera and the camera array by employing spatio-temporal optimization, and produced synthesized views of these poses.

G. Discussion

Due to the particular structure and other specific properties of light fields, there is a growing number of applications in both computer vision and computer graphics that use them to try to solve long-standing problems. By taking advantage of spatio-temporal properties, we can imagine the immense potential of light field video. More related applications, such as tracking and motion capture, can be further exploited. Besides, the integration with more novel imaging techniques is expected to bring new capabilities to light field applications; for instance, by combining light fields with hyperspectral imaging, the extraction of the material properties of the observed object can be greatly enhanced.

VIII. LIGHT FIELD EDITING

Editing light fields has in general been posed as a generalization of traditional 2D image editing. However, given the richer information encoded in the light field, this generalization cannot be considered trivial; as opposed to 2D images, light fields include complex directionally-varying effects such as glossy reflections, refractive objects or occlusions. These problems exist when editing other higher-dimensional media as well, such as video or image collections. We focus on works target to edit the most common structured light fields: We categorize these works on local and global edits, depending on whether they operate on local, precisely placed edits, or on the full light field. Later, we briefly discuss generalization to non-structured light fields and its relation with other types of media.

A. Local Editing

Local point-based editing: One of the first methods for light field editing was introduced by Seitz and Kutulakos [193], [194], based on creating a voxel-based reconstruction of the geometry encoded by the light field. This representation allowed the authors to perform operations such as painting or scissoring. However, the editing is highly dependent on the quality of the voxelization, imposing a Lambertian assumption on the surfaces, which limits its applicability to complex appearances (Fig. 17(a)). Zhang *et al.* [83] proposed a system for performing complex edits in the light field such as erasing, copy-pasting at different depths or cloning, based on patch-based methods. By using a precomputed depth map, the system was able to propagate the edits performed in a single view to the rest of the light field using patch-based synthesis.

Local deformations: Chen *et al.* [197] demonstrated a technique for deforming a light field in the context of photorrealistic image-based rendering, based on splitting the light field into multiple sub-light fields. These were later deformed independently, and then rendered together to keep consistency. To preserve illumination consistency on the deformation, the system is limited to co-axial illumination. Birklbauer *et al.* [198] proposed a method to perform non-linear edits on the spatial domain of the light field, keeping consistency on the angular domain. This allowed complex deformations on the spatial domain of the light field, while preserving light field capabilities such as refocusing.

Editing interaction: The previous works focus on specific editing tools, that require precise positioning of the edit in the three-dimensional scene implicitly encoded by a light field. Jarabo *et al.* [196] analyzed how novice users edited light fields, focusing on the two main positioning paradigms in the literature. These two paradigms used either focus or parallax to specify the position in depth within the light field. Later, Masia *et al.* [199] and Ortin *et al.* [200] further analyzed the workflows of the users during editing. An example edit from [196] can be seen in Fig. 17(c). These works analyzed the interaction using standard 2D displays, although using more complex prototype light field displays (see Sec.tion IX) has been proposed, allowing light field editing by means of gesture tracking [201] or a 3D-positioned light pen [202].

B. Global Editing

Global edits are the most wide-spread form of light field editing, and they have been incorporated in commercial software, usually based on applying some predefined filters (e.g., *Lytro Desktop* [4] or *Lightfield Iris* [203]).

Horn and Chen [204] proposed a shader-based language to manipulate light fields, enabling light field compositing, or complex ray-warping with applications in non-physical defocus. Cossairt *et al.* [205] presented a novel image-based method for compositing real and synthetic objects in the same scene using two 4D light fields, one captured from and one projected onto the real scene. Zhang *et al.* [195] and follow-up work by Wang *et al.* [206] proposed a method for morphing two light fields, based on user-defined correspondences between both light fields (Fig. 17(b)).

Birklbauer and Bimber [207] proposed an approach for light field retargeting, allowing compression or expansion of the spatial domain, while preserving angular consistency without the need of explicit depth estimation. Garces *et al.* [208] recently proposed a method for obtaining the intrinsic decomposition of light fields into albedo and shading. As opposed to other multidimensional intrinsic decomposition methods (e.g. [209], [210]), they leveraged the coherence and structure of the light field.

In the context of coarse edit propagation, Jarabo *et al.* [211] extended *AppProp* [212] to the 4 domain of light fields. To efficiently handle the large scale of light fields, they proposed to operate on a downsampled light field, and then upsample the results following [213]. Several works have followed a similar approach, reparametrizing the light field to increase the locality of the propagation metric [214], or increasing the angular consistency on the propagation distance metric to improve spatio-temporal coherence [215]. A different approach is that based on filtering, by which the editing operation is posed as a multi-dimensional filtering operation over the light field. This is the approach taken by the work of Gryaditskaya *et al.* [216] for gloss editing in light fields.

C. Beyond Structured Light Fields

The works described so far target structured light fields, which have the nice property of a well-defined structure that can be leveraged for robust correspondence estimation for transfer edits (see Section V). This structure is the key difference with other methods targeting non-structured media, such as unstructured light fields [217], [218], video [209], [219]–[221] or image collections [222]–[224]. These are more complex structures, and might require the use of techniques such as depth estimation, optical flow or feature matching, to transfer edits between images.

D. Discussion

While the first works on light field editing date from the late 90's, it has been recently when it has become a more active field with the widespread of plenoptic imaging. In this context, most current works aim to transfer well-established editing tools in traditional 2D images to (structured) light fields (e.g. filters, edit propagation, retargeting or intrinsic decomposition). These works focus on overcoming the intrinsic difficulties of the additional angular domain, instead of leveraging the novel capabilities brought by light fields. This type of editing approaches also often rely on depth information. We have seen how depth from light fields is an active research area (Section V), but a critical question that remains to be answered is how precise this recovered depth information needs to be in order for editing approaches to work well. Of course, this depends on the approach, but should be an aspect addressed by the novel editing approaches to come that rely on depth information.

This research direction is important, and together with principled analyses on the interaction with light fields, should allow for a seamless transition from 2D to 4D images. However, the gained complexity of light fields is still underused in terms of editing capabilities. Making use of all the encoded additional information to enable novel realistic edits on the captured scenes, such as high-quality relighting or material editing, or to increase the expressivity of the artistic editing operations on the final images is a complementary avenue for future research, which would further differentiate between traditional and plenoptic images.

IX. Algorithms for Light Field Display

Displays capable of showing light field contents have been around since the beginning of the 20th Century [225], [226]. Since then, and especially in the last two decades, different technologies exist capable of displaying light fields. In terms of viewing experience, and in comparison to a conventional 2D display or 3D (stereoscopic) display, a light field display adds the ability to see the scene from different viewpoints. This entails that, in light field displays, the perception of depth is conveyed not only through binocular disparity—as in stereoscopic displays—and monocular cues, but also through motion parallax thanks to the different views; this can be done without the need to wear glasses or other hardware.To do this, they must display different views of the light field to both eyes and for each viewpoint within the FOV of the display.

Therelevant work in terms of light field processing for display revolves around two main areas. The first is novel display architectures that require more or less heavy processing of the light field contents prior to its display. The second, a number of

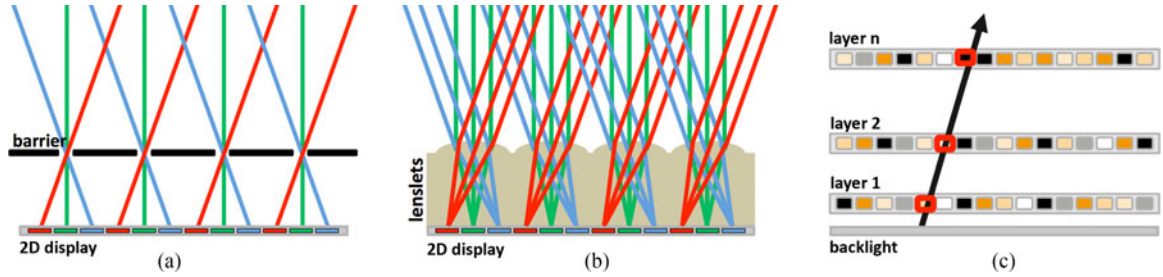


Fig. 18. Three light field display architectures, featuring different ways of creating the final light field rendition. (a) Parallax-barrier based display, in which a mask blocks certain rays from the back layer LCD to show the different views from different directions. (b) Integral imaging display, where a lenslet array is used instead of the mask in (a), also redirecting rays but allowing for a higher light throughput. (c) Example architecture of a multilayer display (specifically, the one presented in [242]); light rays from the backlight go through different attenuation layers, forming the final displayed images. Images adapted from [240].

algorithms for content preprocessing devoted to improving the visual quality of the displayed imagery, attempting to overcome the limitations that the hardware architecture imposes. In the following we review both areas.

A. Light Field Display Architectures

Traditional light field displays: Traditionally, light field displays rely on parallax barriers or lenslet arrays to display these different views to different eyes and viewpoints, e.g., [227]. A back panel displays the different images of the light field, while a front layer redirects light rays appropriately to different positions in front of the screen. This front layer can be formed by regularly-spaced slits which act as a mask in the case of a parallax barrier, or by a series of lenses redirecting rays in the case of a lenslet array-based display (also termed integral imaging); see Fig. 18(a) and (b). In any case, most of these systems require simple image processing, such as some sort of interlacing of the light field views before presentation (see, e.g., [228]) and are out of the scope of this survey. We refer the interested reader to comprehensive existing surveys [229]–[232].

The main drawbacks of these systems are: a low light throughput (in the case of parallax barriers), a trade-off between spatial and angular resolution, and a limited FOV. Additionally, crosstalk between views, inducing ghosting artifacts, is a common problem. In order to alleviate some of these problems, a number of systems have been developed that employ head or eye tracking and actively steer the views accordingly [233]–[235]. However, these approaches require a significant engineering and calibration effort, and typically feature complex systems and/or additional gear. Therefore, more recently, researchers have looked for alternative display architectures, typically at the cost of a heavier processing of the content prior to display.

Multilayer light field displays: The majority of subsequent approaches to displaying light fields have relied on multilayer architectures of one kind or another. A light field is a 4D structure, while a conventional display is 2D. A way of addressing this mismatch is to stack several layers, such as several LCDs. The main goal of the initial multilayer displays was to provide a wider DOF [236]. Further, having a multilayer setup allows to improve accommodation cues, in an attempt to avoid the possible mismatch between focus (accommodation) cues of the eyes, and other depth cues employed by our visual system to infer depth, such as vergence; this mismatch can cause discomfort in the viewer. In these multilayer systems, an algorithm is required to distribute the input light field to the different layers of the display. Often, the distribution is done such that each ray is assigned to the depth layer closest to the ray origin [237]; this of course requires knowledge of scene depth. A more sophisticated approach was presented by Ranieri *et al.* [238], based on decomposing the light field into a series of components which are then rendered to the different display primitives: essentially, different parts, such as diffuse, occlusions, glossy areas, are decomposed and rendered differently. Later works achieved correct accommodation in multi-plane displays [239]; in them, the patterns for the different layers are obtained via primal-dual optimization in the frequency domain. All these approaches require an increase in computation and processing with respect to the traditional ones, but in turn allow to reduce some of their drawbacks. In the following we focus on a family of light field displays which also feature a multilayer architecture, but with a distribution among layers based on a low-rank light field decomposition. This enables them to improve spatial resolution, light throughput and contrast, at the cost of increased processing.

Compressive light field displays: The so-called *compressive* light field displays are multilayer displays consisting of stacked LCD panels (or attenuating layers if content is static) and backlighting (Fig. 18(c)). They can also feature temporal multiplexing of the contents of the different layers (then termed multilayer *multiframe* displays). Prior to display, the content of each of these layers has to be determined, and this is the relevant part in terms of light field processing: the most general formulation can be found in the work on tensor displays [240]; this work aims to provide a general framework for multilayer multiframe displays, encompassing a number of variants which had been appearing before. Lanman *et al.* [241] showed that a low rank approximation of the light field could be used; this, combined with temporal multiplexing, allowed for display of light fields using a dual-layer display. Wetzstein *et al.* [242] posed the problem as a tomographic reconstruction of the light field. In both cases, displaying the light field requires solving a nonnegative least squares problem which yields the optimal attenuation maps to be displayed in the different layers. A generalization to multiple layers using directional backlighting was presented later [240]. The inclusion of time multiplexing [243], and the support of near-correct accommodation cues [244] followed.

These displays recreate the desired scene achieving higher resolution, larger DOF, and higher light throughput than previous approaches, at the cost of significant computation for prior processing of the input light field.In order to reduce the cost of this processing, Heide *et al.* [245] presented an approach that combines adaptive sampling and knowledge of display-specific characteristics to synthesize light fields for compressive displays at a much lower consumption of resources. On a related note, there have also been approaches devoted to fast rendering of light fields from the 3D description of the scene, based on exploiting the redundancy of the light field reusing radiance evaluation between pixels [246].

B. Light Field Display Content Preprocessing

In addition to processing strictly required for display on the different hardware systems, a number of works have strived to improve the quality of the imagery displayed by performing some preprocessing of the content to be shown. These software solutions essentially focus on addressing two of the main limitations that we have mentioned: crosstalk between views; and reduced resolution and the subsequent aliasing, which is particularly severe in the angular domain, in which resolution is most reduced. Additionally, there is a body of work concerned with retargeting of light field content for its display on devices of different characteristics.

Crosstalk: In both parallax barriers and lenslet arrays, light leakage occurs between adjacent views. There is a variety of causes of crosstalk, ranging from imperfect alignment of components to quality of the optical elements, and including the tilted nature of many parallax barriers and lenslet arrays [247]. Crosstalk not only severely diminishes image quality, but can also cause perceived depth to be incorrect [248], [249]. Consequently, a number of works have appeared that are devoted to minimizing the effects of crosstalk in these displays; we focus here on those which imply preprocessing of the input (light field) content. Most approaches incorporate some model of the crosstalk and manipulate the luminance of the input images so that, after being shown on the display with crosstalk, the resulting image is as similar as possible to the target one. To model the crosstalk, careful calibration is often required. Jain and Konrad simulated the crosstalk as a blurring filter, and Wiener deconvolution was employed to preprocess the images to minimize the effect of the crosstalk [250]; the effect of angular antialiasing filters was also taken into account in the process. A linear model of crosstalk was also employed by Wang *et al.* [251], thus modeling the process as a linear system of equations. In their case, the crosstalk coefficient matrix of the model was not measured experimentally, but derived from screen characteristics instead. They further proposed a method to solve the system posing it as a box-constrained integer least squares problem. Li *et al.* [252] focus on light leakage among subpixels: they also rely on inverse filtering for vertical neighboring subpixels, while horizontally blending weights for the neighboring subpixels are computed based on minimizing the mean square error. A different approach is that of Pei *et al.* [253], who propose a method to reduce crosstalk via light field decomposition: they propose to backproject rays to the LCD plane, during which multiple rays can be mapped to one LCD subpixel, these are weighted and the value used to compute the final emitted intensity.

Aliasing: A classical problem of conventional light field displays is spatial and angular *aliasing*. The former has been addressed in seminal works on light field sampling [36], [254]. The latter was tackled by works [255], [256] that derived theoretical formulations of the angular aliasing problem for different cases, and proposed antialiasing techniques based on prefiltering of the content. Some works have also looked into the generation of multiview content from stereoscopic video taking into consideration the angular aliasing problem and the limited depth range of the content that it imposes [88], [257].[1] Image quality offered by these displays is also highly affected by the viewer's position; this was addressed by Du *et al.* [258], who proposed shearing (both globally and locally) the light field, followed by stitching, to hide transitions, improve continuity, and reduce hot-spotting issues. Building on this, a method to eliminate the depth-reversal that takes place in certain viewpoints based on adjusting light field alignment has recently been presented [259]. The problem of hot-spotting has also been addressed by tracking viewers eye positions and dynamically assigning appropriate views to them [260].

Retargeting: Retargeting refers to the process of modifying content to adapt it to the characteristics of a specific display [261]. In the most common case the concept refers to spatial retargeting. The first approach for light fields was that of Birklbauer and Bimber [207], in which they could expand or compress a light field without the need of depth information by applying seam carving on the z-stack ("an image stack that contains only the in-focus contributions in each of its slices while the out-of-focus contributions are removed"). Additionally, a large number of works have addressed the problem of*depth* retargeting. The reason is that the maximum allowed depth budget of a display is limited by the problem of angular aliasing. A traditional solution to angular aliasing, as mentioned before, is pre filtering the content. However, filtering for angular aliasing creates a depth-of-field effect by which only objects close to the screen plane in depth can be shown in focus, while in objects further away only the lower frequencies are preserved and thus appear blurred. To avoid this, depth is retargeted to a narrower depth budget, which essentially requires a subsequent warping of the light field according to the new depth that can be done using techniques targeted to stereo content [262], [263]. Depth retargeting techniques are extensive and not the main focus of this survey; we thus only gloss over them briefly: early works explore simple depth remapping functions [264], [265], while others rely on perceptual considerations [266]–[268], or devise a means of cutting through the EPI volume generating views with multiple centers of projection [269]. A joint image and depth retargeting based on seam carving for light fields has also been proposed [270].

[1]For works tackling the generation of multiview content from a lower dimensional or subsampled input in the general case please refer to Section IV.

C. Discussion

If nowadays stereoscopic displays, aiming at displaying an image plus depth, still suffer from a number of well-known limitations, it is not hard to imagine that light field displays, aiming at displaying a full (4D) light field, suffer even more limitations. When we are attempting to display a 4D structure of visual information on a screen, compromises need to be made. Researchers have resorted to multilayer and multiframe architectures to alleviate the problem. There are two main problems with the current trends. First, the large computational cost is required to calculate the contents of each layer and frame. A light field entails a very large amount of information, and solving the optimizations required for the multilayer decomposition and streaming this information in real time is a challenge. Still, real time implementations exist and this problem may find a solution as computational capability increases and algorithms improve. Second, current displays still fail short of recreating a faithful viewing experience. Before this technology is ready for the market, we need to push farther to solve the limitations we have explained in this section. We need to increase more the FOV, to solve the mismatch in visual cues (e.g., lack of proper accommodation cues), improve energy consumption (increase light efficiency), spatio-angular resolution, and overall image quality. For this purpose, Kovács *et al.* [271] presented a quantitative evaluation for light field displays in terms of spatial and angular resolution. With an industry moving away from stereo displays (in favor of increased spatial resolution or color and luminance fidelity, and with the exception of HMDs), light fields may have the chance of bringing 3D back. However, for this to happen we still need significant improvements in acquisition, compression, and display algorithms for light fields. We have discussed many breakthroughs in these areas in the present survey. Among these, of particular importance for displays are the acquisition or generation of light fields with high enough spatial and angular resolution (Secs. III and IV). Both are needed for a faithful viewing experience, but angular resolution is critical to provide accommodation; it is still unclear and an active research topic how high this resolution needs to be in order to provide correct accommodative cues in the viewer. Content processing for certain types of displays also requires depth information (e.g., some multilayer approaches presented in Section IX-A, or retargeting techniques in Section IX-B); although many are demonstrated with synthetic content featuring perfect depth, techniques such as those in Section V are needed for these approaches to be practical and usable with real footage.

As a side note, with the recent rise in popularity of virtual reality and augmented reality, partly driven by the reduced form factor and cost of current head-mounted displays (HMDs), a new application area is opening up for light fields. On the one hand, they are gaining importance for content generation for these type of devices. On the other, some recent works have used light fields to correct spatial calibration of optical see-through HMDs [272], [273]. In these works, the authors propose a light field-based method to compensate for spatial distortion of images and blur caused by optical see-through HMD optics: the distortion is modeled as a 4D to 4D mapping between light fields, and in the case of blur a PSF model is added (the blur in HMDs is view-dependent).

Finally, it has been shown that light fields and holograms can be seen as complementary representations, and transformations between them have been derived [117].

X. Public Datasets and Code

This section introduces several open source datasets and software tools for light field imaging and manipulation. While this list is not exhaustive, it provides a set of useful links for light field research.

Several light field datasets that include synthetic, real-world and microscopy light fields have been made public. The (New) Stanford light field archive [24] contains thirteen high spatial and angular resolution light fields acquired with a Lego gantry, seven microscope light fields, four light fields acquired with a gantry and two light fields acquired with a camera array. The Stanford Lytro Light Field Archive [274] contains abundant light fields acquired with a Lytro light field camera [4], and is divided into nine different categories. Reřábek and Ebrahimi [275] also provided 118 light fields captured with a Lytro camera [4]. The dataset is divided into ten categories, and can be used for light field compression and quality evaluation. The (old) HCI light field dataset [74] contains synthetic light fields and light fields captured using a Raytrix camera [5]; each synthetic light field includes ground truth depth maps for every viewpoint. Recently, Honauer *et al.* [276] presented a light field benchmark that provides a new HCI light field dataset and an evaluation of depth estimation techniques. Wetzstein [277] provided a number of synthetic light fields containing either 5×5 or 7×7 views. Kim *et al.* [25] provided high spatial and angular resolution light field datasets captured by mounting a consumer DSLR camera on a motorized linear stage. Jarabo *et al.* [278] provided the light fields used in [196], including both synthetic light fields and light fields captured with a Lytro camera. Finally, Ramamoorthi [279] shared a dataset for material recognition using light fields [160].

In terms of available open-sourced software, Dansereau [35] provided a set of tools that allow decoding, calibration, rectification, color correction, basic filtering and visualization of light fields. Wanner *et al.* [66] provided code for light field analysis in terms of both depth estimation and light field super-resolution. Tao *et al.* [102] provided their code for depth estimation that combined defocus cues and correspondence cues. Wang *et al.* [99] and Jeon *et al.* [98] shared their code for light field depth estimation. Kalantari *et al.* [85] shared the source code for learning-based light field view synthesis. Finally, Jarabo *et al.* [196] made public their interface for light field editing.

XI. Conclusion

As an alternative way to capture the visual information in a scene, light fields have become increasingly popular due to their capabilities in the representation of visual information, and the subsequent applications. Dimensionality reduction from the seven dimensions of the plenoptic function to a 4D

representation promoted light fields beyond a theoretical concept. From a hardware perspective, we have witnessed the development of light field acquisition approaches from expensive and bulky camera array systems to portable consumer cameras, such as Raytrix [5], Lytro [4] and PiCam [22]. When light fields were first introduced to computer graphics [3], [11], their applications were only digital refocusing and rendering of virtual views. Today, however, the ability to expeditely capture abundant datasets enables numerous applications, including image enhancement, segmentation, detection, recognition, depth sensing, 3D reconstruction and 3D display.

Meanwhile, there are still many scientific challenges for researchers to address. The inherent resolution trade-off between spatial, angular and temporal resolution is one of the key issues. Although recent works have mitigated the problem by employing spatial, angular or temporal super-resolution techniques, obtaining a high spatial and angular resolution light field with high efficiency still requires more research efforts. Moreover, compared with active depth acquisition approaches, such as structured light and time-of-flight imaging, light field imaging suffers from several problems, such as low data efficiency and high computing complexity. In addition, current consumer plenoptic cameras can deal with only a very limited DOF due to their narrow baseline. Therefore, it is critical to design algorithms that are capable of accurate and efficient depth recovery from light fields with narrow baselines.

In the near future, light field capture based on compact multi-camera arrays will be more and more popular given its high efficiency in getting high resolution data in the spatial, temporal and angular dimensions. Heterogeneous or hybrid camera arrays with different settings of cameras will enrich the diversity of collected information, reduce the data size, and improve the performance of various computer vision algorithms. The pursuit of high performance imaging capabilities will drive novel light field acquisition approaches, such as the estimation of surface or material properties of the scene by combining light fields with hyperspectral imaging. The developments of light field compression and streaming technologies will increase the popularity of light fields in the multimedia industry. The ability of snapshot axial dimension imaging using light field microscopes opens a new door for biology research. Moreover, light field displays are an inevitable trend in VR/AR display systems. The demand of wider FOV, larger DOF and higher dynamic range photography may push research on light field imaging toward a more sophisticated prospect, such as Lytro Cinema [280]. Such high performance light field capture at a single position will help to obtain a high quality reconstruction for a dynamic scene [281] (including geometry, texture, illumination and motion reconstruction), and further visual scene understanding. By keeping these trends in mind it is easy to imagine how light fields may become a part of our daily life in the future.

This paper has summarized the development of light field imaging, in terms of both the processing methods involved, and the challenges that lie ahead. We hope it will help on the evolution of this technology, promoting the development of new exciting applications and changing the way in which we are able to "see" and "understand" the world.

REFERENCES

[1] G. Arun, "The light field," *J. Math. Phys.*, vol. 18, pp. 51–151, 1936.
[2] H. Adelson, Edward and R. Bergen, James, "The plenoptic function and the elements of early vision," *Comput. Models Visual Process.*, MIT Press, 1991, pp. 3–20.
[3] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Annu. Conf. Comput. Graph. Interactive Techn.*, 1996, pp. 31–42.
[4] "Lytro," 2017. [Online]. Available: https://www.lytro.com/
[5] "RayTrix. 3D light field camera technology." 2017. [Online]. Available: http://www.raytrix.de/
[6] "Pelican Imaging." 2016. [Online]. Available: http://www.pelicanimaging.com
[7] M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz, "Light field microscopy," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 924–934, 2006.
[8] M. Levoy, "Light fields and computational imaging," *Computer*, vol. 39, no. 8, pp. 46–55, 2006.
[9] I. Ihrke, J. Restrepo, and L. Mignard-Debise, "Light fields and computational imaging: Briefly revisiting 25 years of research," *Signal Process. Mag.*, vol. 33, no. 5, pp. 59–69, 2016.
[10] G. Wetzstein, I. Ihrke, D. Lanman, and W. Heidrich, "Computational plenoptic imaging," *Comput. Graph. Forum*, vol. 30, no. 8, pp. 2397–2426, 2011.
[11] C. Michael, G. Steven, J. S. Richard, G. Radek, and S. Rick, "The lumigraph," in *Proc. 23rd Annu. Conf. Comput. Graph. Interactive Techn.*, 1996, pp. 43–54.
[12] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *Int. J. Comput. Vis.*, vol. 1, no. 1, pp. 7–55, 1987.
[13] S. E. Chen, "Quicktime VR—An image-based approach to virtual environment navigation," in *Proc. 22nd Annu. Conf. Comput. Graph. Interactive Techn.*, 1995, pp. 29–38.
[14] M. Leonard and B. Gary, "Plenoptic modeling: An image-based rendering system," in *Proc. 22nd Annu. Conf. Comput. Graph. Interactive Techn.*, 1995, pp. 39–46.
[15] J.-X. Chai, X. Tong, S.-C. Chan, and H.-Y. Shum, "Plenoptic sampling," in *Proc. 27nd Annu. Conf. Comput. Graph. Interactive Techn.*, 2000, pp. 307–318.
[16] Z. Lin and H.-Y. Shum, "A geometric analysis of light field rendering," *Int. J. Comput. Vis.*, vol. 58, no. 2, pp. 121–138, 2004.
[17] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen, "Unstructured lumigraph rendering," in *Proc. 28th Annu. Conf. Comput. Graph. Interactive Techn.*, 2001, pp. 425–432.
[18] J. C. Yang, M. Everett, C. Buehler, and L. McMillan, "A real-time distributed light field camera," in *Proc. 13th Eurograph. Workshop Rendering*, 2002, pp. 77–86.
[19] C. Zhang and T. Chen, "A self-reconfigurable camera array," in *Proc. 15th Eurograph. Conf. Symp. Rendering Techn.*, 2004, pp. 243–254.
[20] B. Wilburn *et al.*, "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, 2005.
[21] "ViewPLUS Inc. ProFUSION 25 5x5 camera array system." 2009. [Online]. Available: http://www.viewplus.co.jp/products/profusion25/index-e.html
[22] K. Venkataraman *et al.*, "PiCam: An ultra-thin high performance monolithic camera array," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 2504–2507, 2013.
[23] X. Lin, J. Wu, G. Zheng, and Q. Dai, "Camera array based light field microscopy," *Biomed. Opt. Express*, vol. 6, no. 9, pp. 3179–3189, 2015.
[24] "Stanford (New) Light Field Archive." 2008. [Online]. Available: http://lightfield.stanford.edu/lfs.html
[25] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, 2013, Art. no. 73.
[26] C.-K. Liang, T.-H. Lin, B.-Y. Wong, C. Liu, and H. H. Chen, "Programmable aperture photography: Multiplexed light field acquisition," *ACM Trans. Graph.*, vol. 27, no. 3, 2008. Art. no. 55.
[27] Y. Taguchi, A. Agrawal, S. Ramalingam, and A. Veeraraghavan, "Axial light field for curved mirrors: Reflect your perspective, widen your view," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 499–506.
[28] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Stanford Univ. Comput. Sci., Stanford, CA,USA Tech. Rep, 2005-02, 2005.

[29] D. Lanman, D. Crispell, M. Wachs, and G. Taubin, "Spherical catadioptric arrays: Construction, multi-view geometry, and calibration," in *Proc. Int. Symp. 3D Data Process.*, 2006, pp. 81–88.

[30] T. Georgiev, K. C. Zheng, B. Curless, D. Salesin, S. Nayar, and Chintan,"Spatio-angular resolution tradeoffs in integral photography," in *Proc. Eurograph. Symp. Rendering*, 2006, pp. 263–272.

[31] A. Manakov *et al.*, "A reconfigurable camera add-on for high dynamic range, multispectral, polarization, and light-field imaging," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 47:1–47:14, 2013.

[32] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," in *Proc. SIGGRAPH*, 2007, Art. no. 69.

[33] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–11, 2013.

[34] N. Antipa, S. Necula, R. Ng, and L. Waller, "Single-shot diffuser-encoded light field imaging," in *Proc. IEEE Int. Conf. Comput. Photography*, 2016, pp. 1–11.

[35] D. Dansereau, "Light-field Toolbox for Matlab," 2015. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/49683-light-field-toolbo x-v0-4

[36] A. Isaksen, L. McMillan, and S. J. Gortler, "Dynamically reparameterized light fields," in *Proc. Comput. Graph. Interactive Techn.*, 2000, pp. 297–306.

[37] J. Stewart, J. Yu, S. J. Gortler, and L. McMillan, "A new reconstruction filter for undersampled light fields," in *Proc. Eurograph. Workshop Rendering*, 2003, pp. 150–156.

[38] F. Durand, N. Holzschuch, C. Soler, E. Chan, and F. X. Sillion, "A frequency analysis of light transport," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 1115–1126, 2005.

[39] C. Soler, K. Subr, F. Durand, N. Holzschuch, and F. Sillion, "Fourier depth of field," *ACM Trans. Graph.*, vol. 28, no. 2, 2009, Art. no. 18.

[40] T. Amalia, *Linear Ray and Wave Optics in Phase Space*. Amsterdam, The Netherlands: Elsevier, 2005.

[41] Z. Zhang and M. Levoy, "Wigner distributions and how they relate to the light field," in *Proc. IEEE Int. Conf. Comput. Photography*, 2009, pp. 1–10.

[42] M. Broxton *et al.*, "Wave optics theory and 3-D deconvolution for the light field microscope," *Opt. Express*, vol. 21, no. 21, pp. 25 418–25 439, 2013.

[43] R. Ng, "Fourier slice photography," in *Proc. SIGGRAPH*, 2005, pp. 735–744.

[44] I. Ihrke, G. Wetzstein, and W. Heidrich, "A theory of plenoptic multiplexing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 483–490.

[45] B. Wilburn, M. Smulski, K. Lee, and M. A. Horowitz, "The light field video camera," in *Proc. SPIE Electron. Imag.*, 2002.

[46] B. Wilburn, N. Joshi, V. Vaish, M. Levoy, and M. Horowitz, "High speed video using a dense array of cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. II-294–II-301.

[47] S.-C. Chan, K.-T. Ng, Z.-F. Gan, K.-L. Chan, and H.-Y. Shum, "The plenoptic video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 12, pp. 1650–1659, Dec. 2005.

[48] Y. Liu, Q. Dai, and W. Xu, "A real time interactive dynamic light field transmission system," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2006, pp. 2173–2176.

[49] J. Unger, A. Wenger, T. Hawkins, A. Gardner, and P. Debevec, "Capturing and rendering with incident light fields," in *Proc. Eurograph. Workshop Rendering Techn.*, 2003, pp. 141–149.

[50] J. F. G. W. Donald, G. Dansereau, Glenn Schuster, "A wide-field-of-view monocentric light field camera," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2017, pp. 5048–5057.

[51] I. Ihrke, T. Stich, H. Gottschlich, M. Magnor, and H.-P. Seidel, "Fast incident light field acquisition and rendering," *J. WSCG*, vol. 16, no. 1, pp. 25–32, Feb. 2008.

[52] G. Lippmann, "La photographie intégrale," *Académie des Sci.*, vol. 146, pp. 446–551, 1908.

[53] E. H. Adelson and J. Y. A. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 99–106, Feb. 1992.

[54] L. Wei, C. Liang, G. Myhre, C. Pitts, and K. Akeley, "Improving light field camera sample design with irregularity and aberration," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 152:1–152:11, 2015.

[55] G. Todor and L. Andrew, "Superresolution with plenoptic camera 2.0," Adobe Systems, Inc., San Jose, CA, USA, Tech. Rep, 2009.

[56] R. Horstmeyer, G. Euliss, R. Athale, and M. Levoy, "Flexible multimodal camera using a light field architecture," in *Proc. IEEE Intell. Comput. Commun. Process.*, 2009, pp. 1–8.

[57] M. Levoy, Z. Zhang, and I. McDowall, "Recording and controlling the 4D light field in a microscope using microlens arrays," *J. Microscopy*, vol. 235, pp. 144–162, 2009.

[58] R. Prevedel *et al.*, "Simultaneous whole-animal 3D imaging of neuronal activity using light-field microscopy," *Nature Methods*, vol. 11, no. 7, pp. 727–730, 2014.

[59] T. Nöbauer *et al.*, "Video rate volumetric Ca^{2+} imaging across cortex using seeded iterative demixing (SID) microscopy," *Nature Methods*, vol. 14, pp. 811–818, 2017.

[60] L. Mignard-Debise and I. Ihrke, "Light-field microscopy with a consumer light-field camera," in *Proc. Int. Conf. 3D Vis.*, 2015, pp. 335–343.

[61] D. Lanman, R. Raskar, A. Agrawal, and G. Taubin, "Shield fields: Modeling and capturing 3D occluders," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 15–19, 2008.

[62] A. Ashok and M. A. Neifeld, "Compressive light field imaging," *Proc. SPIE*, vol. 7690, 2010, p. 76900Q.

[63] N. C. Pégard, H.-Y. Liu, N. Antipa, M. Gerlock, H. Adesnik, and L. Waller, "Compressive light-field microscopy for 3D neural activity recording," *Optica*, vol. 3, no. 5, pp. 517–524, 2016.

[64] N. Cohen *et al.*, "Enhancing the performance of the light field microscope using wavefront coding," *Opt. Express*, vol. 22, no. 20, pp. 24 817–24 839, 2014.

[65] Y. Wang, L. Wang, D. Kong, and B. Yin, "High-resolution light field capture with coded aperture," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 5609–5618, Dec. 2015.

[66] S. Wanner and B. Goldlüecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 606–619, Mar. 2014.

[67] V. Boominathan, K. Mitra, and A. Veeraraghavan, "Improving resolution and depth-of-field of light field cameras using a hybrid imaging system," in *Proc. IEEE Int. Conf. Comput. Photography*, 2014, pp. 1–10.

[68] Y. Wang, Y. Liu, W. Heidrich, and Q. Dai, "The light field attachment: Turning a DSLR into a light field camera using a low budget camera ring," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 10, pp. 2357–2364, 2016.

[69] S. Pujades, F. Devernay, and B. Goldluecke, "Bayesian view synthesis and image-based rendering principles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3906–3913.

[70] Z. Zhang, Y. Liu, and Q. Dai, "Light field from micro-baseline image pair," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3800–3809.

[71] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field reconstruction using deep convolutional network on EPI," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, 2017, pp. 6317–6327.

[72] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 24–32.

[73] T.-C. Wang, J.-Y. Zhu, N. K. Kalantari, A. A. Efros, and R. Ramamoorthi, "Light field video capture using a learning-based hybrid imaging system," *ACM Trans. Graph.*, vol. 36, no. 4, Jul. 2017, Art. no. 133.

[74] S. Wanner, S. Meister, and B. Goldlücke, "Datasets and benchmarks for densely sampled 4D light fields," *Vis., Model. Visual.*, pp. 225–226, 2013.

[75] C.-K. Liang and R. Ramamoorthi, "A light transport framework for lenslet light field cameras," *ACM Trans. Graph.*, vol. 34, no. 2, Feb. 2015, Art. no. 16.

[76] W.-S. Chan, E. Y. Lam, M. K. Ng, and G. Y. Mak, "Super-resolution reconstruction in a computational compound-eye imaging system," *Multidimensional Syst. Signal Process.*, vol. 18, no. 2, pp. 83–101, 2007.

[77] T. E. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEETrans. Pattern Anal. Mach. Intell.* vol. 34, no. 5, pp. 972–986, May 2012.

[78] K. Mitra and V. Ashok, "Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior," in *Proc. IEEE Comput. Soc. Conf.Comput. Vis. Pattern Recognit. Workshops*, 2012, pp. 22–28.

[79] R. A. Farrugia, C. Galea, and C. Guillemot, "Super resolution of light field images using linear subspace projection of patch-volumes," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1058–1071, Oct. 2017.

[80] P. James, B. Mike, and D. P. Luigi, "Plenoptic layer-based modeling for image based rendering," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3405–3419, Sep. 2013.

[81] S. Wanner and B. Goldluecke, "Spatial and angular variational super-resolution of 4D light fields," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2012, pp. 608–621.

[82] G. Chaurasia, S. Duchêne, O. Sorkine-Hornung, and G. Drettakis, "Depth synthesis and local warps for plausible image-based navigation," *ACM Trans. Graph.*, vol. 32, 2013, Art. no. 30.

[83] F.-L. Zhang, J. Wang, E. Shechtman, Z.-Y. Zhou, J.-X. Shi, and S.-M. Hu, "Plenopatch: Patch-based plenoptic image manipulation," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 5, pp. 1561–1573, May 2017.

[84] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "Deepstereo: Learning to predict new views from the world's imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5515–5524.

[85] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, 2016, Art. no. 193.

[86] A. Levin and F. Durand, "Linear view synthesis using a dimensionality gap light field prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1831–1838.

[87] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous Fourier domain," *ACM Trans. Graph.*, vol. 34, no. 1, 2014, Art. no. 12.

[88] P. Didyk, P. Sitthi-Amorn, W. Freeman, F. Durand, and W. Matusik, "Joint view expansion and filtering for automultiscopic 3D displays," in *Proc. ACM SIGGRAPH*, 2015, pp. 3906–3913.

[89] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light field reconstruction using Shearlet transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, http://ieeexplore.ieee.org/document/7817742/.

[90] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Accelerated Shearlet-domain light field reconstruction," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1082–1091, Oct. 2017.

[91] D. Mahajan, F.-C. Huang, W. Matusik, R. Ramamoorthi, and P. Belhumeur, "Moving gradients: A path-based method for plausible image interpolation," *ACM Trans. Graph.*, vol. 28, no. 3, 2009, Art. no. 42.

[92] S. Meyer, O. Wang, H. Zimmer, M. Grosse, and A. Sorkine-Hornung, "Phase-based frame interpolation for video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1410–1418.

[93] J. Wu, B. Xiong, X. Lin, J. He, J. Suo, and Q. Dai, "Snapshot hyperspectral volumetric microscopy," *Sci. Rep.*, vol. 6, p. 24624, 2016.

[94] Z. Xiong, L. Wang, H. Li, D. Liu, and F. Wu, "Snapshot hyperspectral light field imaging," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3270–3278.

[95] H. Rueda, C. Fu, D. L. Lau, and G. R. Arce, "Spectral-TOF compressive snapshot camera: Towards hyperspectral+depth imagery," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 992–1003, Oct. 2017.

[96] Z. Yu, X. Guo, H. Ling, A. Lumsdaine, and J. Yu, "Line assisted light field triangulation and stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2792–2799.

[97] S. Heber and T. Pock, "Shape from light field meets robust PCA," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2014, pp. 751–767.

[98] H. G. Jeon, J. Park, G. Choe, and J. Park, "Accurate depth map estimation from a lenslet light field camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1547–1555.

[99] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3487–3495.

[100] H. Zhu, Q. Wang, and J. Yu, "Occlusion-model guided anti-occlusion depth estimation in light field," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 965–978, Oct. 2017.

[101] C. Chen, H. Lin, Z. Yu, S. B. Kang, and J. Yu, "Light field stereo matching using bilateral statistics of surface cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1518–1525.

[102] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 673–680.

[103] M. W. Tao, P. P. Srinivasan, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi, "Depth from shading, defocus, and correspondence using light-field angular coherence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1940–1948.

[104] J. Tian, Z. Murez, T. Cui, Z. Zhang, D. Kriegman, and R. Ramamoorthi, "Depth and image restoration from light field in a scattering medium," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1–10.

[105] W. Williem and I. K. Park, "Robust light field depth estimation for noisy scene with occlusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4396–4404.

[106] H. Lin, C. Chen, S. B. Kang, and J. Yu, "Depth recovery from light field using focal stack symmetry," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3451–3459.

[107] J. Y. Lee and R.-H. Park, "Depth estimation from light field by accumulating binary maps based on the foreground and background separation," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 955–964, Oct. 2017.

[108] M. Tao, J.-C. Su, T.-C. Wang, J. Malik, and R. Ramamoorthi, "Depth estimation and specular removal for glossy surfaces using point and line consistency with light-field cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1155–1169, Jun. 2016.

[109] T.-C. Wang, M. Chandraker, A. Efros, and R. Ramamoorthi, "SVBRDF-invariant shape and reflectance estimation from light-field cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5451–5459.

[110] J. Li, M. Lu, and Z.-N. Li, "Continuous depth map reconstruction from light fields," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3257–3265, Nov. 2015.

[111] B. Krolla, M. Diebold, B. Goldlücke, and D. Stricker, "Spherical light fields," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–12.

[112] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4D light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 41–48.

[113] M. Diebold, B. Jähne, and A. Gatto, "Heterogeneous light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1745–1753.

[114] S. Heber and T. Pock, "Convolutional networks for shape from light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3746–3754.

[115] A. Criminisi, S. B. Kang, R. Swaminathan, S. Richard, and P. Anandan, "Epipolar-plane image analysis: An approach to determining structure from motion," *Comput. Vis. Image Understanding*, vol. 97, no. 1, pp. 51–85, 2005.

[116] T. Basha, S. Avidan, A. Sorkine-Hornung, and W. Matusik, "Structure and motion from scene registration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1426–1433.

[117] R. Ziegler, S. Bucheli, L. Ahrenberg, M. Marcus, and G. Markus, "A bidirectional light field—Hologram transform," *Comput. Graph. Forum*, vol. 26, no. 3, pp. 435–446, 2007.

[118] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Comput. Vis. Image Understanding*, vol. 145, pp. 148–159, 2016.

[119] O. Jonhannsen, A. Sulc, and B. Goldluecke, "What sparse light field coding reveals about scene structure," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3262–3270.

[120] S. Heber, R. Ranftl, and T. Pock, "Variational shape from light field," in *Proc. Int. Conf. Energy Minimization Methods Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 66–79.

[121] M. Diebold *et al.*, "Light-field camera design for high-accuracy depth estimation," *Proc. SPIE*, vol. 9528, 2015, Art. no. 952803.

[122] "ISO 12233," Photography-electronic still picture cameras-resolution measurements, 2000.

[123] Q. Fu, Z. Zhou, Y. Yuan, and B. Xiangli, "Image quality evaluation of light field photography," *Image Quality Syst. Performance*, vol. 7867, no. 4, pp. 357–366, 2011.

[124] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[125] H. Shidanshidi, F. Safaei, and W. Li, "A quantitative approach for comparison and evaluation of light field rendering techniques," in *Proc. IEEE Int. Conf. Multimedia Expo.*, vol. 50, no. 4, 2011, pp. 1–4.

[126] V. K. Adhikarla *et al.*, "Towards a quality metric for dense light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 58–67.

[127] A. Aggoun, "A 3D DCT compression algorithm for omnidirectional integral images," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2006, pp. 517–520.

[128] A. Aggoun, "Compression of 3D integral images using 3D wavelet transform," *J. Display Technol.*, vol. 7, no. 11, pp. 586–592, 2011.

[129] M. Magnor, A. Endmann, and B. Girod, "Progressive compression and rendering of light fields," in *Proc. Vis., Model. Visual.*, 2000, pp. 199–204.

[130] D. Xu, Q. Dai, and W. Xu, "Data compression of light field using wavelet packet," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2004, pp. 1071–1074.

[131] C.-L. Chang, X. Zhu, P. Ramanathan, and B. Girod, "Light field compression using disparity-compensated lifting and shape adaptation," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 793–806, Apr. 2006.

[132] M. Magnor and B. Girod, "Data compression for light-field rendering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 3, pp. 338C–343, Apr. 2000.

[133] C. Conti, P. Kovacs, T. Balogh, P. Nunes, and L. D. Soares, "Light-field video coding using geometry-based disparity compensation," in *Proc. 3DTV-Conf., The True Vision - Capture, Transmiss. Display 3D Video*, 2014, pp. 1–4.

[134] S. Kundu, "Light field compression using homography and 2D warping," in *Proc. IEEE Int. Conf. Acoust.*, 2012, pp. 1349–1352.

[135] X. Jiang, M. L. Pendu, R. Farrugia, and C. Guillemot, "Light field compression with homography-based low rank approximation," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1132–1145, Oct. 2017.

[136] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*, 2016, pp. 1–4.

[137] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Scalable coding of plenoptic images by using a sparse set and disparities," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 1, pp. 80–91, Jan. 2016.

[138] X. Jin, H. Han, and Q. Dai, "Image reshaping for efficient compression of plenoptic content," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1173–1186, Oct. 2017.

[139] F. Dai, J. Zhang, Y. Ma, and Y. Zhang, "Lenselet image compression scheme based on subaperture images streaming," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 4733–4737.

[140] A. Vieira, H. Duarte, C. Perra, L. Tavora, and P. Assuncao, "Data formats for high efficiency coding of lytro-illum light fields," in *Proc. Int. Conf. Image Process. Theory*, 2015, pp. 494–497.

[141] L. Li, Z. Li, B. Li, D. Liu, and H. Li, "Pseudo sequence based 2-D hierarchical coding structure for light-field image compression," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, Oct. 2017.

[142] C. Perra, "Lossless plenoptic image compression using adaptive block differential prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 1231–1234.

[143] P. Helin, P. Astola, B. Rao, and I. Tabus, "Sparse modelling and predictive coding of subaperture images for lossless plenoptic image compression," in *Proc. 3DTV-Conf.: The True Vision - Capture, Transmiss. Display 3D Video*, 2016, pp. 1–4.

[144] P. Helin, P. Astola, B. Rao, and I. Tabus, "Minimum description length sparse modeling and region merging for lossless plenoptic image compression," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1146–1161, Oct. 2017.

[145] "JPEG Pleno," 2017. [Online]. Available: https://jpeg.org/jpegpleno/index.html

[146] "ICME 2016 Grand Challenge on Light-Field Image Compression." 2016. [Online]. Available: http://mmspg.epfl.ch/ICME2016GrandChallenge_1

[147] C. Liang, "Predictive light field compression," Jul. 21 2016, U.S. Patent 15/085 048. [Online]. Available: https://www.google.com/patents/US20160212443

[148] I. Viola, M. Řeřábek, and T. Ebrahimi, "Comparison and evaluation of light field image coding approaches," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1092–1106, Oct. 2017.

[149] M. Levoy, B. Chen, V. Vaish, M. Horowitz, I. McDowall, and M. Bolas, "Synthetic aperture confocal imaging," *ACM Trans. Graph.*, vol. 23, pp. 825–834, 2004.

[150] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Linear volumetric focus for light field cameras," *ACM Trans. Graph.*, vol. 34, no. 2, pp. 15:1–15:20, 2015.

[151] J. T. Kajiya *et al.*, "The rendering equation," *ACM Trans. Graph.*, vol. 20, pp. 143–150, 1986.

[152] L. Yatziv, G. Sapiro, and M. Levoy, "Lightfield completion," in *Proc. IEEE Int. Conf. Image Process.*, 2004, pp. 1787–1790.

[153] V. Vaish, R. Szeliski, C. L. Zitnick, S. B. Kang, and M. Levoy, "Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 2331–2338.

[154] T. Yang *et al.*, "All-in-focus synthetic aperture imaging," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2014, pp. 1–15.

[155] N. Joshi, S. Avidan, W. Matusik, and K. D. J., "Synthetic aperture tracking: Tracking through occlusions," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[156] T. Yang, Y. Zhang, and X. Tong, "A new hybrid synthetic aperture imaging model for tracking and seeing people through occlusion in," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 9, pp. 1461–1475, Sep. 2013.

[157] S. Wanner, C. Straehle, and B. Goldluecke, "Globally consistent multi-label assignment on the ray space of 4D light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1011–1018.

[158] K. Yücer, A. Sorkine-Hornung, O. Wang, and O. Sorkine-Hornung, "Efficient 3D object segmentation from densely sampled light fields with applications to 3D reconstruction," *ACM Trans. Graph.*, vol. 35, no. 3, 2016, Art. no. 22.

[159] K. Yücer, C. Kim, A. Sorkine-Hornung, and O. Sorkine-Hornung, "Depth from gradients in dense light fields for object reconstruction," in *Proc. Int. Conf. 3D Vis.*, 2016, pp. 249–257.

[160] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4D light-field dataset and CNN architectures for material recognition," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2016, pp. 121–138.

[161] N. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Automatic 3D object segmentation in multiple views using volumetric graph-cuts," *Image Vis. Comput.*, vol. 28, no. 1, pp. 14–25, 2010.

[162] N. D. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Automatic object segmentation from calibrated images," in *Proc. Conf. Vis. Media Production*, 2011, pp. 126–137.

[163] M. Hog, N. Sabater, and C. Guillemot, "Light field segmentation using a ray-based graph structure," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2016, pp. 35–50.

[164] J. Berent and D. Luigi, "Segmentation of epipolar-plane image volumes with occlusion and disocclusion competition," in *Proc. IEEE Workshop Multimedia Signal Process.*, 2006, pp. 182–185.

[165] H. Zhu, Q. Zhang, and Q. Wang, "4D light field superpixel and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6384–6392.

[166] O. Johannsen, A. Sulc, and B. Goldluecke, "Variational separation of light field layers," in *Proc. Vis., Model. Vis.*, 2015, pp. 135–142.

[167] Y. Xu, H. Nagahara, A. Shimada, and R. Taniguchi, "Transcut: Transparent object segmentation from a light-field image," in *Proc. IEEE Int. Conf. Comput. Vis*, 2015, pp. 3442–3450.

[168] N. Joshi, W. Matusik, and S. Avidan, "Natural video matting using camera arrays," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 779–786, 2006.

[169] D. Cho, S. Kim, and Y.-W. Tai, "Consistent matting for light field images," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2014, pp. 90–104.

[170] J. Fiss, B. Curless, and R. Szeliski, "Light field layer matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 623–631.

[171] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2806–2813.

[172] Z. Pei, Y. Zhang, T. Yang, X. Zhang, and Y.-H. Yang, "A novel multi-object detection method in complex scene using synthetic aperture imaging," *Pattern Recognit.*, vol. 45, pp. 1637–1658, 2012.

[173] R. Gross, I. Matthews, and S. Baker, "Eigen light-fields and face recognition across pose," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, vol. 44, no. 41, pp. 3–9.

[174] R. Gross, I. Matthews, and S. Baker, "Appearance-based face recognition and light-fields," *IEEE Trans. Pattern Aanl. Mach. Intell.*, vol. 26, no. 4, pp. 449–465, Apr. 2004.

[175] R. Raghavendra, B. Yang, K. B. Raja, and C. Busch, "A new perspective—face recognition with light-field camera," in *Proc. Int. Conf. Biometrics*, 2013, pp. 1–8.

[176] R. Raghavendra, B. Yang, K. B. Raja, and C. Busch, "Robust iris recognition using light-field camera," in *Proc. Colour Visual Comput. Symp.*, 2013, pp. 1–6.

[177] G. Wetzstein, D. Roodnick, R. Raskar, and W. Heidrich, "Refractive shape from light field distortion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1180–1186.

[178] G. Wetzstein, R. Raskar, and W. Heidrich, "Hand-held Schlieren photography with light field probes," in *Proc. IEEE Int. Conf. Comput. Photography*, 2011, pp. 1–8.

[179] K. Maeno, H. Nagahara, A. Shimada, and R.-I. Taniguichi, "Light field distortion feature for transparent object recognition," in *Proc IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2786–2793.

[180] P. Anglin, S. J. Reeves, and B. S. Thurow, "Characterization of plenoptic imaging systems and efficient volumetric estimation from plenoptic data," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1020–1033, Oct. 2017.

[181] Y. Ding, F. Li, Y. Ji, and J. Yu, "Dynamic fluid surface acquisition using a camera array," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2478–2485.

[182] A. Sulc, A. Alperovich, N. Marniok, and B. Goldluecke, "Reflection separation in light fields based on sparse coding and specular flow," in *Proc. Conf. Vis., Model. Vis.*, 2016, pp. 137–144.

[183] F. Lu, L. He, S. You, and Z. Hao, "Identifying surface BRDF from a single 4D light field image via deep neural network," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1047–1057, Oct. 2017.

[184] D. G. Dansereau, I. Mahon, O. Pizarro, and S. B. Williams, "Plenoptic flow: Closed-form visual odometry for light field cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2011, pp. 4455–4462.

[185] N. Zeller, F. Quint, and U. Stilla, "From the calibration of a light-field camera to direct plenoptic odometry," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1004–1019, Oct. 2017.

[186] I. Schillebeeckx and R. Pless, "Structured light field design for correspondence free rotation estimation," in *Proc. IEEE Int. Conf. Comput. Photography*, 2015, pp. 1–10.

[187] P. P. Srinivasan, M. W. Tao, R. Ng, and R. Ramamoorthi, "Oriented light-field windows for scene flow," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3496–3504.

[188] T. Fahringer, K. Lynch, and B. S. Thurow, "Volumetric particle image velocimetry with a single plenoptic camera," *Meas. Sci. Technol.*, vol. 26, no. 11, pp. 115–201, 2015.

[189] X. Guo, Z. Yu, S. B. Kang, H. Lin, and J. Yu, "Enhancing light fields through ray-space stitching," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 7, pp. 1852–1861, Jul. 2016.

[190] O. Johannsen, A. Sulc, and B. Goldluecke, "On linear structure from motion for light field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 720–728.

[191] P. P. Srinivasan, R. Ng, and R. Ramamoorthi, "Light field blind motion deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3958–3966.

[192] B. M. Smith, L. Zhang, H. Jin, and A. Agarwala, "Light field video stabilization," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 341–348.

[193] S. M. Seitz and K. N. Kutulakos, "Plenoptic image editing," in *Proc. IEEE 6th Int. Conf. Comput. Vis.*, 1998, pp. 17–24.

[194] S. M. Seitz and K. N. Kutulakos, "Plenoptic image editing," *Int. J. Comput. Vis.*, vol. 48, no. 2, pp. 115–129, 2002.

[195] Z. Zhang, L. Wang, B. Guo, and H.-Y. Shum, "Feature-based light field morphing," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 457–464, 2002.

[196] A. Jarabo, B. Masia, A. Bousseau, F. Pellacini, and D. Gutierrez, "How do people edit light fields?" *ACM Trans. Graph.*, vol. 33, no. 4, pp. 70–79, 2014.

[197] B. Chen, E. Ofek, H.-Y. Shum, and L. M., "Interactive deformation of light fields," in *Proc. Symp. Interactive 3D Graph. Games*, 2005, pp. 139–146.

[198] C. Birklbauer, D. C. Schedl, and O. Bimber, "Nonuniform spatial deformation of light fields by locally linear transformations," *ACM Trans. Graph.*, vol. 35, no. 5, 2016, Art. no. 156.

[199] B. Masia, A. Jarabo, and D. Gutierrez, "Favored workflows in light field editing," in *Proc. Comput. Graph., Vis., Comput. Vis. Image Process.*, 2014, pp. 1–8.

[200] M. Ortin, A. Jarabo, B. Masia, and D. Gutierrez, "Analyzing interfaces and workflows for light field editing," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1162–1172, Oct. 2017.

[201] F. Marton, M. Agus, E. Gobbetti, and M. B. Rodriguez, "Natural exploration of 3D massive models on large-scale light field displays using the fox proximal navigation technique," *Comput. Graph.*, vol. 36, no. 8, pp. 893–903, 2012.

[202] J. Tompkin *et al.*, "Interactive light field painting," *ACM SIGGRAPH 2012 Emerg. Technol.*, 2012.

[203] Vertical Horizon, "Lightfield Iris," 2015. [Online]. Available: http://www.verticalhorizon-software.com/LFI/index.html

[204] R. Horn, Daniel and B. Chen, "Lightshop: Interactive light field manipulation and rendering," in *Proc. Symp. Interactive 3D Graph.*, 2007, pp. 121–128.

[205] O. Cossairt, S. K. Nayar, and R. Ramamoorthi, "Light field transfer: Global illumination between real and synthetic objects," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 57:1–57:6, 2008.

[206] L. Wang, S. Lin, S. Lee, B. Guo, and Y. Shum, Heung, "Light field morphing using 2D features," *IEEE Trans. Vis. Comput. Graph.*, vol. 11, no. 1, pp. 25–34, Jan./Feb. 2005.

[207] C. Birklbauer and O. Bimber, "Light-field retargeting," *Comput. Graph. Forum*, vol. 31, 2012, pp. 295–303.

[208] E. Garces, J. I. Echevarria, W. Zhang, H. Wu, K. Zhou, and D. Gutierrez, "Intrinsic light field images," *Comput. Graph. Forum*, Wiley Online Library, 2017, pp. 1–10.

[209] G. Ye, E. Garces, Y. Liu, Q. Dai, and D. Gutierrez, "Intrinsic video and applications," *ACM Trans. Graph.*, vol. 33, no. 4, 2014, Art. no. 80.

[210] S. Duchêne *et al.*, "Multi-view intrinsic images of outdoors scenes with an application to relighting," *ACM Trans. Graph.*, vol. 34, 2015, Art. no. 164.

[211] A. Jarabo, B. Masia, and D. Gutierrez, "Efficient propagation of light field edits," in *Proc. Ibero-Amer. Symp. Comput. Graph.*, 2011, pp. 75–80.

[212] X. An and F. Pellacini, "Appprop: All-pairs appearance-space edit propagation," *ACM Trans. Graph.*, vol. 27, pp. 40:1–40:9, 2008.

[213] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, 2007, Article no. 96.

[214] H. Ao *et al.*, "Light field editing based on reparameterization," in *Proc. Pacific Rim Conf. Multimedia*, 2015, pp. 601–610.

[215] Williem, K. W. Shon, and I. K. Park, "Spatio-angular consistent editing framework for 4D light field images," *Multimedia Tools Appl.*, vol. 75, no. 23, 2016, pp. 16615–16631.

[216] Y. Gryaditskaya, B. Masia, P. Didyk, K. Myszkowski, and H.-P. Seidel, "Gloss editing in light fields," in *Proc. Int. Workshop Vis., Model. Vis.*, 2016, pp. 127–135.

[217] A. Davis, M. Levoy, and F. Durand, "Unstructured light fields," *Comput. Graph. Forum*, vol. 31, no. 2, pp. 305–314, 2012.

[218] A. Wender, J. Iseringhausen, B. Goldluecke, M. Fuchs, and M. B. Hullin, "Light field imaging through household optics," in *Proc. Vis., Model. Vis.*, 2015, pp. 159–166.

[219] K. Xu, Y. Li, T. Ju, S.-M. Hu, and T.-Q. Liu, "Efficient affinity-based edit propagation using k-d tree," *ACM Trans. Graph.*, vol. 28, pp. 118:1–118:6, 2009.

[220] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, "A system for retargeting of streaming video," *ACM Trans. Graph.*, vol. 28, no. 5, 2009, Art. no. 126.

[221] A. Meka, M. Zollhöfer, C. Richardt, and C. Theobalt, "Live intrinsic video," *ACM Trans. Graph.*, vol. 35, no. 4, 2016, Art. no. 109.

[222] S. W. Hasinoff, M. Jozwiak, F. Durand, and W. T. Freeman, "Search-and-replace editing for personal photo collections," in *Proc. IEEE Int. Conf. Comput. Photograph.*, 2010, pp. 1–8.

[223] Y. HaCohen, E. Shechtman, D. B. Goldman, and D. Lischinski, "Non-rigid dense correspondence with applications for image enhancement," *ACM Trans. Graph.*, vol. 30, no. 4, 2011, Art. no. 70.

[224] K. Yücer, A. Jacobson, A. Hornung, and O. Sorkine, "Transfusive image manipulationin," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 176:1–176:9, 2012.

[225] F. E. Ives, "Parallax stereogram and process of making same," U.S. Patent 725 567, Apr. 14, 1903.

[226] G. Lippmann, "Epreuves reversibles donnant la sensation du relief," *J. Phys. Theor. Appl.*, vol. 7, no. 1, pp. 821–825, 1908.

[227] W. Matusik and H. Pfister, "3D TV: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 814–824, 2004.

[228] K. Perlin, S. Paxia, and J. S. Kollin, "An autostereoscopic display," in *Proc. 27th Annu. Conf. Comput. Graph. Interactive Techn.*, 2000, pp. 319–326.

[229] H. Urey, K. V. Chellappan, E. Erden, and P. Surman, "State of the art in stereoscopic and autostereoscopic displays," *Proc. IEEE*, vol. 99, no. 4, pp. 540–555, Apr. 2011.

[230] E. Lueder, *3D Displays*. New York, NY, USA: Wiley, 2011, vol. 32.

[231] G. Wetzstein, D. Lanman, D. Gutierrez, and M. Hirsch, "Computational displays," in *Proc. ACM SIGGRAPH Course Notes*, 2012.

[232] B. Masia, G. Wetzstein, P. Didyk, and D. Gutierrez, "A survey on computational displays: Pushing the boundaries of optics, computation, and perception," *Comput. Graph.*, vol. 37, no. 8, pp. 1012–1038, 2013.

[233] G. Woodgate, D. Ezra, N. Holliman, B. Omar, R. Moseley, and J. Harrold, "Autostereoscopic display and method of controlling an autostereoscopic display," Sep. 15, 1998, U.S. Patent 5 808 792. [Online]. Available: https://www.google.com/patents/US5808792

[234] T. Peterka, R. L. Kooima, D. J. Sandin, A. Johnson, J. Leigh, and T. A. DeFanti, "Advances in the dynallax solid-state dynamic parallax barrier autostereoscopic visualization display system," *IEEE Trans. Vis.Comput. Graph.*, vol. 14, no. 3, pp. 487–499, May/Jun. 2008.

[235] S.-K. Kim, K.-H. Yoon, S. K. Yoon, and H. Ju, "Parallax barrier engineering for image quality improvement in an autostereoscopic 3D display," *Opt. Express*, vol. 23, no. 10, pp. 13 230–13 244, 2015.

[236] H. Gotoda, "A multilayer liquid crystal display for autostereoscopic 3D viewing," in *Proc. IS&T/SPIE Electron. Imag.*, 2010, pp. 75 240–75 240.

[237] N. Ranieri *et al.*, "Multi-layered automultiscopic displays," *Comput. Graph. Forum*, vol. 31, no. 7, pp. 2135–2143, 2012.
[238] N. Ranieri, S. Heinzle, P. Barnum, W. Matusik, and M. Gross, "32.3: Light-field approximation using basic display layer primitives," in *Proc. SID Symp. Digest. Techn. Papers*, vol. 44, no. 1, pp. 408–411, 2013.
[239] R. Narain, R. A. Albert, A. Bulbul, G. J. Ward, M. S. Banks, and J. F. O'Brien, "Optimal presentation of imagery with focus cues on multi-plane displays," *ACM Trans. Graph.*, vol. 34, no. 4, 2015, Art. no. 59.
[240] G. Wetzstein, D. Lanman, M. Hirsch, and R. Raskar, "Tensor displays: compressive light field synthesis using multilayer displays with directional backlighting," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 80:1–80:11, 2012.
[241] D. Lanman, M. Hirsch, Y. Kim, and R. Raskar, "Content-adaptive parallax barriers: Optimizing dual-layer 3D displays using low-rank light field factorization," *ACM Trans. Graph.*, vol. 29, no. 6, 2010, Art. no. 163.
[242] G. Wetzstein, D. Lanman, W. Heidrich, and R. Raskar, "Layered 3D: tomographic image synthesis for attenuation-based light field and high dynamic range displays," *ACM Trans. Graph.*, vol. 30, no. 4, 2011, Art. no. 95.
[243] F. Heide, D. Lanman, D. Reddy, J. Kautz, K. Pulli, and D. Luebke, "Cascaded displays: Spatiotemporal superresolution using offset pixel layers," *ACM Trans. Graph.*, vol. 33, no. 4, 2014, Art. no. 60.
[244] A. Maimone, G. Wetzstein, M. Hirsch, D. Lanman, R. Raskar, and H. Fuchs, "Focus 3D: Compressive accommodation display," *ACM Trans. Graph.*, vol. 32, no. 5, pp. 1–13, 2013.
[245] F. Heide, G. Wetzstein, R. Raskar, and W. Heidrich, "Adaptive image synthesis for compressive displays," *ACM Trans. Graph.*, vol. 32, no. 4, 2013, Article no. 132.
[246] A. Jarabo, R. Buisan, and D. Gutierrez, "Bidirectional clustering for scalable VPL-based global illumination," in *Proc. CEIG '15*, 2015, pp. 1–9.
[247] A. J. Woods, "Crosstalk in stereoscopic displays: A review," *J. Electron. Imag.*, vol. 21, no. 4, 2012, Art. no. 040902.
[248] I. Tsirlin, L. M. Wilcox, and R. S. Allison, "The effect of crosstalk on the perceived depth from disparity and monocular occlusions," *IEEE Trans. Broadcasting*, vol. 57, no. 2, pp. 445–453, Jun. 2011.
[249] I. Tsirlin, L. M. Wilcox, and R. S. Allison, "Effect of crosstalk on depth magnitude in thin structures," *J. Electron. Imag.*, vol. 21, no. 1, 2012, Art. no. 011003.
[250] A. Jain and J. Konrad, "Crosstalk in automultiscopic 3-D displays: Blessing in disguise?," *Electron. Imag.*, vol. 6490, 2007, Art. no. 649012.
[251] X. Wang and C. Hou, "Improved crosstalk reduction on multiview 3D display by using BILS algorithm," *J. Appl. Math.*, vol. 2014, 2014, Art. no. 428602.
[252] D. Li, D. Zang, X. Qiao, L. Wang, and M. Zhang, "3D synthesis and crosstalk reduction for lenticular autostereoscopic displays," *J. Display Technol.*, vol. 11, no. 11, pp. 939–946, 2015.
[253] R. Pei, Z. Geng, Z. X. Zhang, X. Cao, and R. Wang, "A novel optimization method for lenticular 3-D display based on light field decomposion," *J. Display Technol.*, vol. 12, no. 7, pp. 727–735, Jul. 2016.
[254] J. Stewart, J. Yu, S. J. Gortler, and L. McMillan, "A new reconstruction filter for undersampled light fields," *ACM Int. Conf. Proc. Series*, vol. 44, pp. 150–156, 2003.
[255] M. Zwicker, W. Matusik, F. Durand, H. Pfister, and C. Forlines, "Antialiasing for automultiscopic 3D displays," in *Proc. ACM SIGGRAPH*, 2006, Article no. 107.
[256] J. Konrad and P. Agniel, "Subsampling models and anti-alias filters for 3-D automultiscopic displays," *IEEE Trans. Image Process.*, vol. 15, no. 1, pp. 128–140, Jan. 2006.
[257] A. Chapiro *et al.*, "Optimizing stereo-to-multiview conversion for autostereoscopic displays," *Comput. Graph. Forum*, vol. 33, no. 2, 2014, pp. 63–72.
[258] S.-P. Du, P. Didyk, F. Durand, S.-M. Hu, and W. Matusik, "Improving visual quality of view transitions in automultiscopic displays," *ACM Trans. Graph.*, vol. 33, no. 6, 2014, Art. no. 192.
[259] S. Xie *et al.*, "Depth-reversal-free three-dimensional display by adjusting the light field alignment," *Opt. Commun.*, vol. 387, pp. 415–420, 2016.
[260] J. Liu, T. Malzbender, S. Qin, B. Zhang, C.-A. Wu, and J. Davis, "Dynamic mapping for multiview autostereoscopic displays," in *SPIE/IS&T Electronic Imag.*, vol. 9391, 2015, Art. no. 9391.
[261] F. Banterle *et al.*, "Multidimensional image retargeting," in *Proc. SIGGRAPH Asia 2011 Courses*, 2011, Article no. 15.
[262] P. Didyk, T. Ritschel, E. Eisemann, K. Myszkowski, and H.-P. Seidel, "Adaptive image-space stereo view synthesis," in *Proc. Vis. Modeling Vis. Workshop*, 2010, pp. 299–306.
[263] N. Plath, S. Knorr, L. Goldmann, and T. Sikora, "Adaptive image warping for hole prevention in 3D view synthesis," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3420–3432, Sep. 2013.
[264] M. Kim, S. Lee, C. Choi, G.-M. Um, N. Hur, and J. Kim, "Depth scaling of multiview images for automultiscopic 3D monitors," in *Proc. 2008 3DTV Conf., True Vis. Capture, Transmiss. Display 3D Video*, 2008, pp. 181–184.
[265] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3D," *ACM Trans. Graph.*, vol. 29, no. 4, 2010, Art. no. 75.
[266] P. Didyk, T. Ritschel, E. Eisemann, K. Myszkowski, and H.-P. Seidel, "A perceptual model for disparity," *ACM Trans. Graph.*, vol. 30, no. 4, 2011, Article no. 96.
[267] P. Didyk, T. Ritschel, E. Eisemann, K. Myszkowski, H.-P. Seidel, and W. Matusik, "A luminance-contrast-aware disparity model and applications," *ACM Trans. Graph.*, vol. 31, no. 6, 2012, Article no. 184.
[268] B. Masia, G. Wetzstein, C. Aliaga, R. Raskar, and D. Gutierrez, "Display adaptive 3D content remapping," *Comput. Graph.*, vol. 37, no. 8, pp. 983–996, 2013.
[269] C. Kim, A. Hornung, S. Heinzle, W. Matusik, and M. Gross, "Multi-perspective stereoscopy from light fields," *ACM Trans. Graph.*, vol. 30, no. 6, 2011, Art. no. 190.
[270] V. Ramachandra, M. Zwicker, and T. Q. Nguyen, "Combined image plus depth seam carving for multiview 3D images," in *Proc. 2009 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 737–740.
[271] P. T. Kovács, R. Bregović, A. Boev, A. Barsi, and A. Gotchev, "Quantifying spatial and angular resolution of light field 3D displays," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, Oct. 2017.
[272] Y. Itoh and G. Klinker, "Light-field correction for spatial calibration of optical see-through head-mounted displays," *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 4, pp. 471–480, Apr. 2015.
[273] Y. Itoh, T. Amano, D. Iwai, and G. Klinker, "Gaussian light field: Estimation of viewpoint-dependent blur for optical see-through head-mounted displays," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 11, pp. 2368–2376, Nov. 2016.
[274] "Stanford Lytro Light Field Archive." 2017. [Online]. Available: http://lightfields.stanford.edu/
[275] M. Řeřábek and T. Ebrahimi, "New light field image dataset," in *Proc. 8th Int.. Conf. Quality Multimedia Exp.*, 2016, pp. 1–2.
[276] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldlücke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 19–34.
[277] "Synthetic Light Field Archive." 2013. [Online]. Available: http://web.media.mit.edu/~gordonw/SyntheticLightFields/
[278] "How do people edit light fields?: Dataset." 2014. [Online]. Available: http://giga.cps.unizar.es/ ajarabo/pubs/lfeiSIG14/data/index.html
[279] "Light Field Datasets." 2017. [Online]. Available: http://cseweb.ucsd.edu/~viscomp/projects/LF/
[280] "Lytro Cinema." 2017. [Online]. Available: https://lytro.com/cinema
[281] K. Guo, F. Xu, Y. Wang, Y. Liu, and Q. Dai, "Robust non-rigid motion tracking and surface reconstruction using L_0 regularization," *IEEE Trans. Visual. Comput. Graph.*, http://ieeexplore.ieee.org/document/7888591/



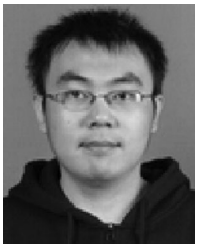
Gaochang Wu received the B.S. and M.S. degrees in mechanical engineering from Northeastern University, Shenyang, China, in 2013, and 2015, respectively. He is currently working toward the Ph.D. degree in the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, China. His current research interests include image processing, machine learning, and computational photography.



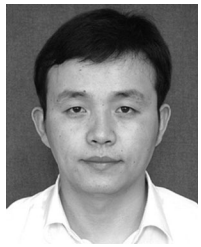
Belen Masia received the Ph.D. degree in computer science from Universidad de Zaragoza, Zaragoza, Spain, in 2013. She has been a Postdoctoral Researcher at the Max Planck Institute for Informatics in Saarbruecken, Germany, and a Visiting Researcher at the Massachusetts Institute of Technology. He is currently an Assistant Professor in the Department of Computer Science, Universidad de Zaragoza, and a member of the I3A Research Institute. Her research in computational imaging, displays, and applied perception has been published in top venues such as ACM Transactions on Graphics. She is an Associate Editor of ACM Transactions on Applied Perception, and of Computers & Graphics. She has also served in numerous technical papers program committees, including ACM SIGGRAPH Asia, Eurographics, and Pacific Graphics. She has received a number of awards, including the Eurographics Young Researcher Award in 2017, a Eurographics PhD Award in 2015, she was selected as one of the top ten Innovators Under 35 in Spain by MIT Technology Review in 2014, and has also received a NVIDIA Graduate Fellowship in 2012. He is a Eurographics Young Fellow.



Adrian Jarabo received the Ph.D. degree in computer graphics from Universidad de Zaragoza, Zaragoza, Spain, in 2015. He is currently a Postdoctoral Researcher at Universidad de Zaragoza. Before that, he was a Visiting Researcher in Trinity College Dublin and Microsoft Research Asia. His research interests include topics such as light transport simulation, appearance modeling, or plenoptic manipulation. He received the Eurographics Ph.D. Award in 2017.



Yuchen Zhang received the B.S. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2013, where he is currently working toward the Ph.D. degree. His main research interests include computational photography and graph signal processing.



Liangyong Wang received the B.E. degree in automation, the M.Sc. and Ph.D. degrees in control theory and control engineering from Northeastern University, Shenyang, China, respectively in 2001, 2004, and 2010. He is currently with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, China. His research interests are cyber-physical systems, machine learning and big data focused on industrial applications.



Qionghai Dai received the B.S. degree in mathematics from Shanxi Normal University, Xi'an, China, in 1987, and the M.E. and Ph.D. degrees in computer science and automation from Northeastern University, China, in 1994 and 1996, respectively. Since 1997, he has been with the faculty of Tsinghua University, Beijing, China, and is currently a Professor and the Director of the Broadband Networks and Digital Media Laboratory. His research areas include video communication, computer vision, and graphics.



Tianyou Chai received the Ph.D. degree in control theory and engineering from Northeastern University, Shenyang, China, in 1985. He has been with the Research Center of Automation, Northeastern University, Shenyang, China, since 1985, where he became a Professor in 1988 and a Chair Professor in 2004. His current research interests include adaptive control, intelligent decoupling control, integrated plant control and systems, and the development of control technologies with applications to various industrial processes. He is a member of the Chinese Academy of Engineering, an Academician of International Eurasian Academy of Sciences, and IFAC Fellow. He is a distinguished visiting fellow of The Royal Academy of Engineering (UK) and an Invitation Fellow of Japan Society for the Promotion of Science.



Yebin Liu received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2002, and the Ph.D. degree from the Automation Department, Tsinghua University, Beijing, China, in 2009. He has been working as a Research Fellow in the Computer Graphics Group, Max Planck Institute for Informatik, Germany, in 2010. He is currently an Associate Professor in Tsinghua University. His research areas include computer vision and computer graphics.

Depth Estimation From Light Field by Accumulating Binary Maps Based on Foreground–Background Separation

Jae Young Lee and Rae-Hong Park, *Senior Member, IEEE*

***Abstract*—A three-dimensional (3-D) scene can be separated into two regions: foreground and background, which are regions in front of and behind the focused plane, respectively. From the 4-D light field, this paper proposes a depth estimation method by accumulating binary maps, which are computed by the foreground–background separation with light field reparameterization. In the proposed foreground–background separation, an optical phenomenon is used where bundles of rays from the background are flipped on their conjugate planes. Using the Lambertian assumption and gradient constraint, the foreground and background of a scene can be converted to a binary map by voting the gradient signs in every angular patch. Using light field reparameterization, the disparity map can be obtained by accumulating the binary maps. Finding the extremum index in the existing methods corresponds to finding the zero crossing index in the proposed method. By accumulating the binary maps, the proposed method has an advantage on the computational efficiency in terms of memory usage, and can use various reparameterization strategies suitable for applications. Experimental results with synthetic and real images show that the proposed method can generate a high-quality disparity map.**

***Index Terms*—Depth estimation, foreground-background separation, light fields, light field re-parameterization.**

I. Introduction

DEPTH maps are widely used to improve the performance of computer vision applications such as object recognition [1], segmentation and pose estimation [2] with a stereo camera or an RGB-D camera. Recently, a computational camera has been developed to acquire additional depth information, which cannot be captured by a conventional camera [3]. A light field camera is a computational camera that is highlighted as a next-generation camera [4]. The light field camera can capture incoming rays with directional information in the four-dimensional (4-D) manifold [5]. Although scene geometry is obviously three-dimensional (3-D) space, a light field camera encodes incoming rays in a 4-D manifold with two-dimensional (2-D) spatial information and 2-D angular information [6]. Many image processing and computer vision researches on light field [6], such as saliency detection [7] and occlusion detection [8], have been presented. Especially, depth estimation using a light field camera plays an important role in research stream and has been increasingly investigated [9]–[16].

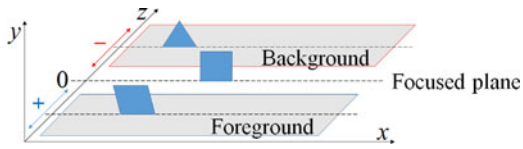


Fig. 1. Definition of the foreground and background in this paper.

Fig. 1 shows the definition of the foreground and background used in this paper, where they can be separated with respect to the focused plane. Using the light field re-parameterization, which is part of the digital refocusing algorithm [17], the focused plane of a 4-D light field image can be moved to the foreground or background. The binary maps that indicate the foreground or background are computed with the re-parameterized light field images. By accumulating the resulting binary maps, disparity map can be obtained. In addition, while applications, such as saliency detection and occlusion detection, are based on a 3-D focal stack [7] or a pre-computed depth information [8], a slice of the proposed cost volume can be directly used in a 4-D light field manifold without a pre-computed depth map as in [18].

The foreground-background separation can be explained by an optical phenomenon [18] that bundles of rays from the background are flipped on the conjugate plane [19]. This phenomenon is also observed on the conventional camera, but only the magnitude of the blurriness is captured [20], whereas the light field camera can capture the direction of the rays [19]. By using the binary maps of the foreground and background, the labeling problems in the existing methods by finding the extremum index in the constructed cost volumes correspond to finding the zero crossing index in the proposed method.

The proposed method has an advantage on the computational efficiency in terms of the memory usage. The existing methods [11]–[13] store the computed cost in a 3-D volume format on the memory, which is called cost volume. In contrast, the proposed method needs to save only the 2-D accumulated results. In addition, using the coarse-to-fine re-parameterization strategy in the cost volume construction step, the proposed method can reduce the waiting time to display outline of the disparity map.

Manuscript received December 27, 2016; revised April 17, 2017, July 15, 2017, and August 21, 2017; accepted August 24, 2017. Date of publication August 30, 2017; date of current version October 23, 2017. This work was supported in part by Brain Korea 21 Plus. The guest editor coordinating the review of this paper and approving it for publication was Prof. Yebin Liu. *(Corresponding author: Rae-Hong Park.)*

The authors are with the Department of Electronic Engineering, Sogang University, Seoul 100–611, South Korea (e-mail: mcneato@sogang.ac.kr; rhpark@sogang.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2747154

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Experimental results show that the proposed method is effective for both synthetic and real light field images. In the stereo taxonomy [21], the proposed method is a local approach while the existing methods are hybrid methods by combining the local and global approaches. Although it might be fair to compare the resulting disparity maps of the proposed method with the initial disparity maps of the existing methods, they are compared with the final disparity maps of the existing methods. Similar to the existing methods, the proposed method also has failure cases in non-Lambertian surfaces and repeated high-frequency regions due to the Lambertian assumption of the proposed method and angular aliasing, respectively. According to the overview paper [6], the proposed method is categorized in the matching-based method, where the foreground -background separation is used as the main feature.

The foreground-background separation was proposed in our previous work [18]. This paper is an extended version of [18] with extensive discussions, analysis, and experiments in view of the depth estimation. The disparity sign voting is appropriately improved for the depth estimation by computing gradients using Sobel filters in the lenslet and sub-aperture images. In addition, the re-parameterization strategies are introduced, and used in the framework of the proposed method.

II. Related Work

A. Overview of the Existing Methods

Unlike stereo vision methods, the existing depth estimation methods from light field generally use both local and global methods in view of the taxonomy of stereo algorithms [21]. According to this taxonomy, the existing depth estimation methods from light field can be generally divided into two steps: initial depth estimation step (local step) and optimization step (global step). The existing initial depth estimation methods can be classified into three groups. The first group uses depth from stereo (DFS) [9], [10]. The second group combines two or more cues: DFS, depth from focus (DFF), or depth from defocus (DFD) [11]–[13]. The third group infers depth information from the epipolar plane images (EPIs) [14]–[16]. The first group uses DFS by measuring the consistency in angular patches. Chen *et al*. [9] proposed a bilateral consistency metric (BCM) to reduce occlusion boundary artifacts. When the disparity of the occluded area is computed, they used the BCM in each angular patch to reduce the influence of foreground (or background) pixels in the occluding region on the background (or foreground) pixels. Jeon *et al*. [10] used the phase shift method for sub-pixel accuracy in stereo matching. They constructed the cost volume using linear combination of the sum of absolute differences and the sum of gradient differences in angular patches. The second group uses two or more initial depths to obtain a high-quality depth map. Tao *et al*. [11] proposed a framework that combines the correspondence and defocus cues. They used the variance in angular patches for the correspondence cue and the sharpness measure in refocused images for the defocus cue. The initial depths from both cues are combined by the Markov random field (MRF). Wang *et al*. [12] modeled occlusion in light field and constructed the cost volume by dividing angular patches at the occlusion boundary into two regions: foreground and background. With the constructed cost volume and edge map, the occlusion map is predicted. The depth map is finally estimated using the MRF. Williem and Park [13] proposed a noise-robust cost volume construction method using the correspondence and defocus cues. The third group measures the slope levels in EPIs. Structure tensor is used in vertical and horizontal EPIs to estimate the initial depth map [14]. Johannsen *et al*. [15] used the patch-based sparse coding in the EPIs. By constructing a depth-aware dictionary, the initial depth map is estimated. Dansereau and Bruton [16] estimated the depth map by using a weighted average of gradient magnitudes in EPIs as weights.

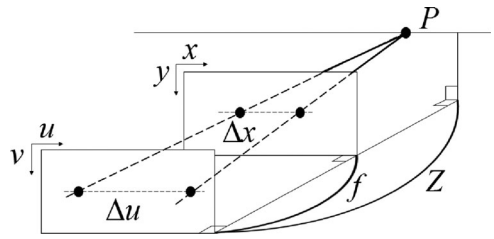


Fig. 2. Light field geometry.

According to the taxonomy of light field depth estimation algorithms [22], the proposed method is regarded as SCam and MultiView without occlusion-aware data term. The proposed method uses the linear interpolation in the re-parameterization step and full views as input. The cost volume is constructed with the constant step size, and therefore, the numbers of depth labels are different depending on the disparity range used. The cost volume is refined by the guided image filter. Finally, the winner-takes-all manner is used for the final depth map.

B. Light Field Re-Parameterization

A light field is mathematically modeled by a 4-D function, $L(x, y, u, v)$, which uses two-plane parameterization [23]: xy plane (spatial domain) and uv plane (angular domain). Each ray is represented as a line by two points that are intersected with xy and uv planes.

Fig. 2 shows the light field geometry with an arbitrary object point P. Using similar triangles, the displacement Δx between adjacent views with the baseline Δu can be expressed as $\Delta x = (f/Z)\Delta u$, where Z represents the object depth and f signifies the distance between the xy and uv planes [14]. Fig. 2 is represented without consideration of different local image coordinates of views. If they are considered, the centers of local image coordinates of different views are aligned. The Δx can be expressed as $\Delta u - \Delta x$ and finally, $\Delta x = (f/Z)\Delta u$ can be derived using similar triangles [14]. Similarly, the relationship between Δy and Δv can be represented as $\Delta y = (f/Z)\Delta v$. Using Lambertian assumption, disparity d, which implies relative displacement, can be expressed as [24]

$$d = \frac{f}{Z} = \frac{\Delta x}{\Delta u} = \frac{\Delta y}{\Delta v}. \tag{1}$$

Using (1), an arbitrary view (u, v) can be expressed in terms of the reference view (u, v) as [24]

$$L_0(x, y, u, v) = L_0(x - (u - u_r)d, y - (v - v_r)d, u_r, v_r), \tag{2}$$

where L_0 denotes the captured light field. By letting $(u_r, v_r) = (0, 0)$, the light field re-parameterization is defined as

$$L_d(x, y, u, v) = L_0(x + ud, y + vd, u, v), \tag{3}$$

where L_d represents a light field re-parameterized by the disparity d. A re-parameterized light field L_d is used to separate the foreground and background by voting the gradient signs in angular patches. While a light field is re-parameterized in terms of the depth in the existing methods, the proposed method re-parameterizes the light field in terms of the disparity based on gradient constraint.

C. Disparity Representation Based on Gradient Constraint

A pair of two adjacent images in terms of sub-aperture images can be considered as a stereo image pair. Thus, it is possible to use the gradient constraint to each pair of adjacent images in the light field [18]. The spatial gradient $\partial L(x, u)/\partial x$ in the xu plane of the light field is linearly approximated as

$$\begin{aligned}\frac{\partial L(x,u)}{\partial x} &\approx \frac{L(x+\Delta x, u) - L(x,u)}{\Delta x} \\ &= \frac{L(x+d\Delta u, u) - L(x,u)}{d\Delta u} \\ &= \frac{L(x, u+\Delta u) - L(x,u)}{d\Delta u} \\ &= \frac{1}{d}\frac{L(x, u+\Delta u) - L(x,u)}{\Delta u},\end{aligned} \tag{4}$$

where Δu represents the baseline between adjacent images. Using (1) and (4), the disparity $\hat{d}$ can be expressed as

$$\hat{d}_{xu} = \frac{(L(x, u+\Delta u) - L(x,u))/\Delta u}{\partial L(x,u)/\partial x} = \frac{\partial L(x,u)/\partial u}{\partial L(x,u)/\partial x}. \tag{5}$$

Similarly, the disparity $\hat{d}_{yv}$ can be defined in the yv plane.

III. Proposed Depth Estimation Method

Fig. 3 shows the block diagram of the proposed method. As shown in Fig. 3(a), an input light field image $L_0(x, y, u, v)$ is re-parameterized with the given d value. With the re-parameterized light field $L_d(x, y, u, v)$, the foreground and background are separated. By accumulating the binary maps $BW(x, y, d)$, a final disparity map $D(x, y)$ is obtained. The main algorithm of the proposed method is the foreground-background separation, which enables to obtain the disparity map by simple accumulation. Fig. 3(b) shows the sub-block diagram of the foreground-background separation. The disparity sign voting sub-block is derived from an optical phenomenon where bundles of rays from the background are flipped. For the disparity, we use this optical phenomenon to distinguish the foreground and background with the relaxed Lambertian assumption, which only checks whether disparity signs computed from the xu and yv planes are the same or not. By voting the computed disparity signs in every angular patch with the pixels satisfying the relaxed Lambertian assumption, $C_V(x, y, d)$ can be obtained. Different from our previous work [18], the foreground-background separation in this paper does not use the gradient magnitude as the weight in the disparity sign voting. The voting result $C_V(x, y, d)$ is aggregated using an edge-preserving filter [25]. The binary maps $BW(x, y, d)$ are obtained by thresholding the edge-preserving smoothing result $\tilde{C}_V(x, y, d)$. The obtained disparity map can be easily converted into the depth map using the pinhole camera model.

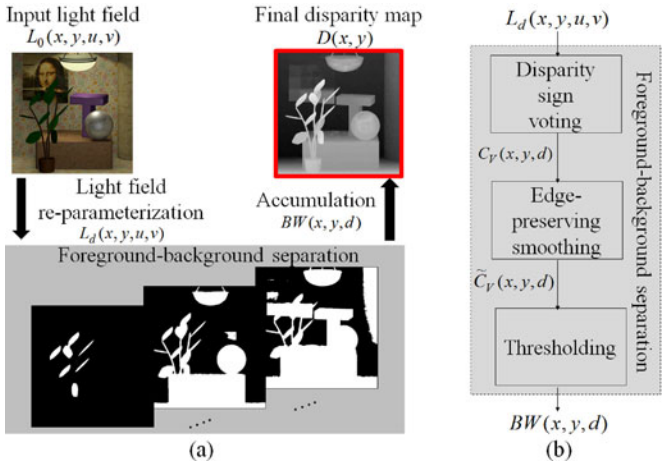


Fig. 3. Block diagram of the proposed method (Mona image in dataset [29]). (a) Overall system of the proposed method. (b) Sub-block diagram of the proposed foreground-background separation.

A. Foreground-Background Separation

Using (5), the disparity signs S_d^{xu} and S_d^{yv} of $\hat{d}_{xu}$ and $\hat{d}_{yv}$ from the xu and yv planes can be expressed as [18]

$$S_d^{xu}(x, y, u, v) = \mathrm{sgn}\left(\left(\frac{\partial L_d}{\partial x}\right)\left(\frac{\partial L_d}{\partial u}\right)\right) = \mathrm{sgn}\left(\hat{d}_{xu}\right) \tag{6}$$

$$S_d^{yv}(x, y, u, v) = \mathrm{sgn}\left(\left(\frac{\partial L_d}{\partial y}\right)\left(\frac{\partial L_d}{\partial v}\right)\right) = \mathrm{sgn}\left(\hat{d}_{yv}\right), \tag{7}$$

respectively, where $\mathrm{sgn}(\cdot)$ represents the sign function. The disparity sign information of (5) can be equivalently expressed as (6) and (7). Similar to the local binary pattern (LBP) [26], the proposed method focuses on the disparity sign rather than disparity magnitude.

The disparity signs are voted in an angular patch by

$$C_V(x, y, d) = \begin{cases} \dfrac{\sum_{u,v}\delta_d(x,y,u,v)\sum_i S_d^i}{2\sum_{u,v}\delta_d(x,y,u,v)}, & \text{if } \sum_{u,v}\delta_d(x,y,u,v) \neq 0; \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where $i = \{xu, yv\}$. $\delta_d(x, y, u, v)$ is the indicator function satisfying the Lambertian assumption using only gradient sign

information. The indicator function $\delta_d(x, y, u, v)$ is defined as

$$\delta_d(x, y, u, v) = \begin{cases} 1, & \text{if } S_d^{xu}(x, y, u, v) = S_d^{yv}(x, y, u, v); \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

to indicate the pixels satisfying the Lambertian assumption by checking the disparity signs from the xu and yv planes. If disparity signs S_d^{xu} and S_d^{yv} are not the same, the corresponding pixels are considered as unreliable. The foreground and background correspond to the positive and negative disparity signs, respectively. Given that a light field image is prone to noise and disparities from the xu and yv planes are rarely the same, the relaxed version of the Lambertian assumption is used. As AdaBoost algorithm [27] combines several weak classifiers to obtain a strong classifier, the results of the foreground-background separation are computed from each stereo pair and combined using a voting algorithm with the indicator function $\delta(x, y, u, v)$. That is, $C_V(x, y, d)$ can be considered as a strong classifier and $S_d^i(x, y, u, v)$ at fixed viewpoints (u, v) can be viewed as weak classifiers.

By aggregating each slice of $C_V(x, y, d)$ using a guided image filter, $\tilde{C}_V(x, y, d)$ is computed. The guided image filter rejects outliers and performs edge-preserving smoothing in $C_V(x, y, d)$. $\tilde{C}_V(x, y, d)$ is thresholded by 0 to separate the foreground and background.

Fig. 4 shows the intermediate results and further analysis of the cost volume of the proposed method using Platonic image [28]. Fig. 4(a) shows the reference view image. In Fig. 4(b), the proposed method is compared with our previous work [18] by showing binary maps obtained with two different disparity values. Two sets of results are obtained by thresholding the ground truth disparity map and re-parameterized input light field with threshold of –0.94 and 0.52, respectively. In each set, from left to right, the ground truth, and results by the proposed method and our previous work are shown, respectively. In our previous work, the edge information is redundantly used in the disparity sign voting and the guided image filtering. However, the proposed method does not use the gradient magnitude as the weight of the gradient sign in the disparity sign voting. Comparison of the binary maps shows that the proposed method gives better foreground-background separation results.

For analysis of the occlusion boundary, in Fig. 4(c) and (d), xd slices of $C_V(x, y, d)$ and $\tilde{C}_V(x, y, d)$ along the line indicated in Fig. 4(a) are shown, respectively, at the top, whereas the cropped slices are shown at the bottom. For easy comparison of $BW(x, y, d)$ and the ground truth, the ground truth is shown in 3-D volume format by thresholding each disparity value. In Fig. 4(e), xd slices of the ground truth 3-D volume and $BW(x, y, d)$ are overlapped, with two small boxes enlarged at the left and right sides. While the estimated disparity signs are generally well-separated with edges preserved using the guided image filter, there are also errors, where red and green pixels represent incorrectly separated foreground and background by the proposed method, respectively. Fig. 4(f) shows the cost profiles at neighboring pixels of boxes indicated in Fig. 4(d) of the proposed method, the existing method (correspondence response) [11], and ground truth near the occlusion boundary,

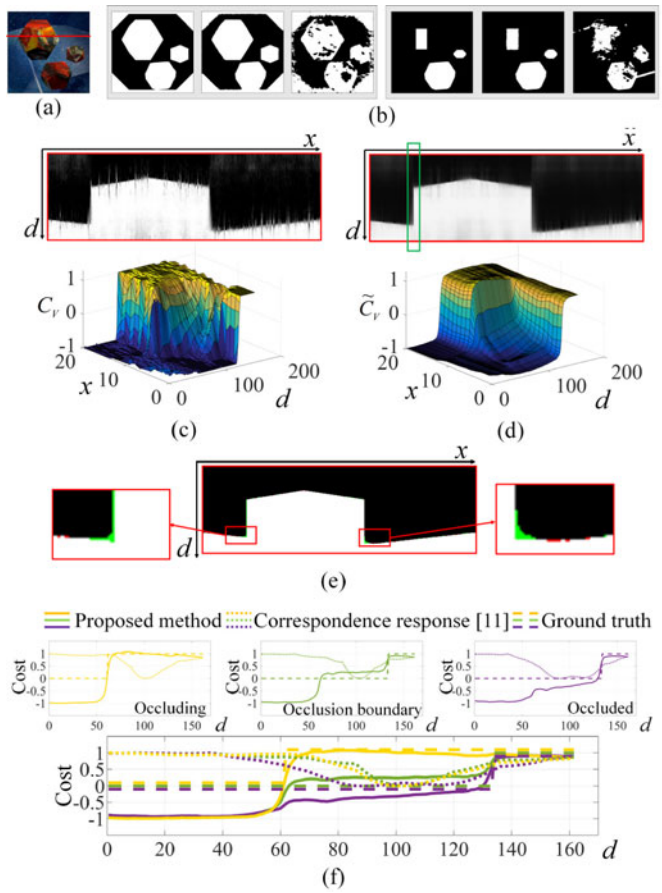


Fig. 4. Intermediate results and further analysis of the cost volume of the proposed method. (a) A reference view of Platonic image in [28]. (b) Binary maps of the ground truth, and the results by the proposed method and our previous work [18] with two different disparity values. (c) xd slice of $C_V(x, y, d)$. (d) xd slice of $\tilde{C}_V(x, y, d)$. (e) Overlapping xd slices of the ground truth and $BW(x, y, d)$.(f) Cost profiles of the proposed method, correspondence response [11], and the ground truth.

respectively. The cost profiles of pixels are shown separately at the top while overlapped at the bottom. At the occlusion boundary, the cost profiles of the proposed method and the correspondence response are affected by the occluding (yellow lines) and occluded regions (purple lines). While the correspondence response has a minimum value between the ground truth disparity labels of occluding and occluded pixels, the proposed method has two steep slopes of occluding and occluded pixels. Since the proposed method determines the disparity labels of thresholded cost profiles by either of two steep slopes, reasonable performance at occlusion boundary can be achieved even though there exists an ambiguous region with the cost value close to zero between two steep slopes.

B. Computation of Light Field Image Gradients

To compute a light field image gradient for each direction, Sobel filters are used. Fig. 5 shows the intermediate results of the foreground-background separation. Fig. 5(a) and (b) show the sub-aperture and lenslet images of Mona image in [29], respectively. The sub-aperture images can be viewed as a set of images from different views, whereas the lenslet images can be

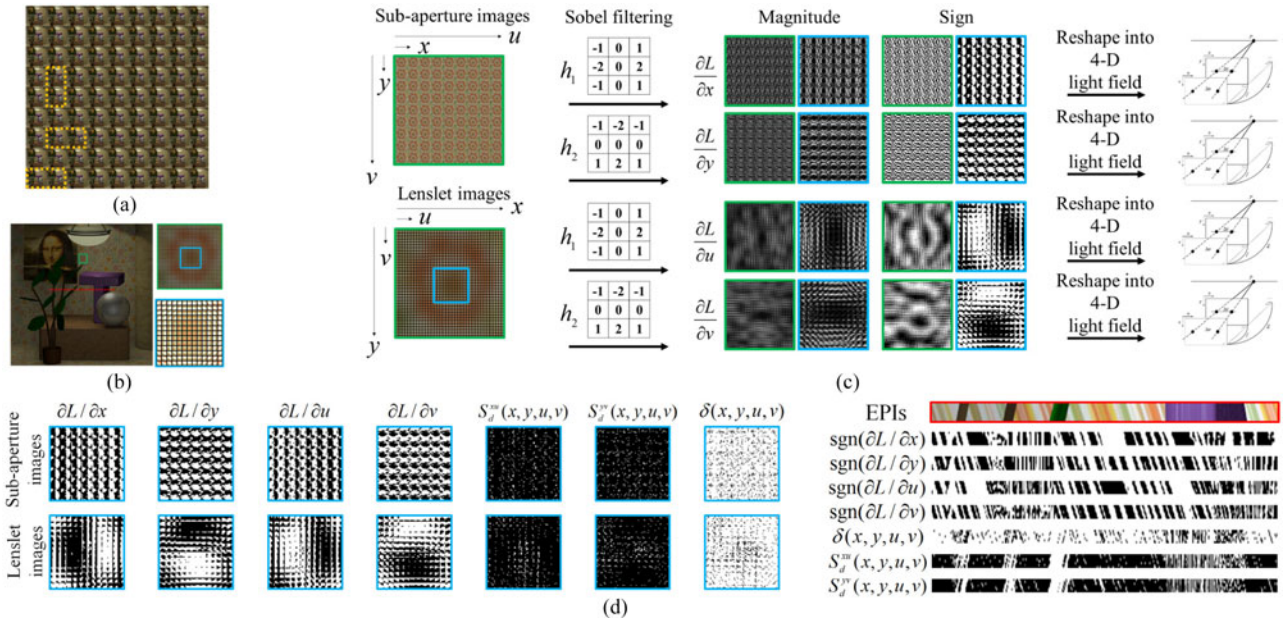


Fig. 5. Intermediate results to obtain the foreground-background separation with Mona image in dataset [29]. (a) Sub-aperture images. (b) Lenslet images. (c) Process of light field image gradient computation. (d) Visualization of intermediate results as sub-aperture, lenslet, and epipolar plane images.

considered as a set of rays decomposed for each direction. The small solid box in Fig. 5(b) is enlarged at the top on the right side of the lenslet image. A box in the enlarged box is further magnified at the bottom of the enlarged box. Fig. 5(c) shows the process of computing the light field image gradient. Sobel filters in h_1 and h_2 are used on the sub-aperture and lenslet images to obtain the spatial gradients ($\partial L/\partial x$, $\partial L/\partial y$) and angular gradients ($\partial L/\partial v$), respectively. Before Sobel filtering is performed, the sub-aperture and lenslet images are zero-padded to prevent influence of neighboring views or lenslets, respectively.

The intermediate results of the foreground-background separation (gradient magnitude and sign) in Fig. 5(c) are shown using a 4-D patch of size $u \times v \times x \times y = 9 \times 9 \times 13 \times 13$ in the input light field image. The gradient magnitudes are represented by using absolute values of the gradient. In Fig. 5(d), the spatial and angular gradients are obtained in the sub-aperture and lenslet images (on the left), respectively, and they are represented as sub-aperture, lenslet, and epipolar plane images for easy comparison (on the right). After computing the light field gradients in the sub-aperture and lenslet images, each gradient map is reshaped into the 4-D light field manifold. S_d^{xu}, S_d^{yv}, and $\delta(x, y, u, v)$ are also represented as sub-aperture and lenslet images although they are computed on the 4-D manifolds. As shown in S_d^{xu}, S_d^{yv}, and $\delta(x, y, u, v)$, the 4-D patch is dominantly classified as negative (background).

In terms of sub-aperture images, two neighboring views can be considered as a stereo pair. The dotted boxes in Fig. 5(a) show three examples of a stereo pair. In terms of lenslet images, each pixel in an angular patch is obtained from different view. That is, neighboring pixels in an angular patch in terms of the lenslet images can be viewed as certain pixels from multiple stereo pairs. Thus, the absolute values of angular gradients of the light field image represent the relationship between neighboring views. In terms of the stereo, the magnitudes of the angular gradients can be viewed as the SAD from multiple stereo pairs. To obtain the depth information in the disparity sign voting, the proposed method aggregates the sign information instead of magnitude, which is derived from relationships mentioned above.

The intermediate results of the dotted line in Fig. 5(b) are represented as the EPIs at the right side of Fig. 5(d). The sign maps of each gradient are shown in the second to fifth images of Fig. 5(d). The indicator function $\delta(x, y, u, v)$ is shown in the sixth row on the right side. The disparity sign maps S_d^{xu} and S_d^{yv} are shown in the last two rows, respectively. The region with different S_d^{xu} and S_d^{yv} signs has the black pixel value and does not contribute to the foreground-background separation in (8). The slope sign of the background is different from that of the foreground, because of the optical phenomenon that the background pixels are flipped in view of the lenslet images. The proposed method uses the relationships between sub-aperture images for initial depth estimation in terms of the foreground-background separation, although the existing methods measure the consistency or the correspondence in each lenslet image in view of stereo taxonomy.

Dansereau and Bruton [16] used Sobel filters for the depth estimation. While the depth estimation is the main goal of both the proposed method and their method, the purpose of Sobel filtering is different. They directly estimated the depth information from the slope of EPIs by performing Sobel filtering on EPIs, while the proposed method performs Sobel filtering to obtain the light field image gradient.

C. Relation to Existing Cost Volume Construction Methods

The existing depth estimation methods construct the cost volume in the first step in their frameworks for estimating the initial depth map [9]–[13]. The existing method based on DFS

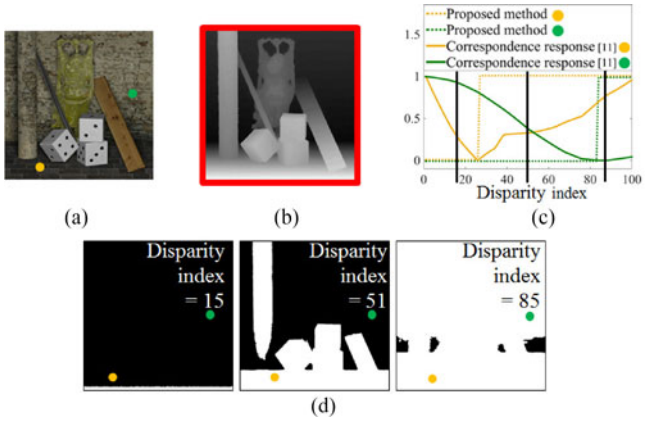


Fig. 6. Relationship between the proposed method and the existing cost volume construction method [11]. (a) Reference view image of Buddha image from [29]. (b) Final disparity map of the proposed method. (c) Profiles of the proposed method and the correspondence response at each dot. (d) Three examples of the proposed method with different disparity indices.

constructs the cost volume by measuring the consistency in every angular patch. By finding the minimum index on cost profiles at each pixel location, the initial depth map can be estimated. The cost volume construction method based on DFF measures the sharpness from the refocused images. By finding the maximum index, the initial depth map from the cost volume based on DFF can be estimated. Using one or more initial depth maps, the final depth is computed based on the optimization techniques such as MRF and graph-cut.

Considering the existing cost volume construction method, the proposed method can be viewed as the initial depth estimation method by finding the zero crossing index in cost profiles at each pixel. Fig. 6 shows the relationship between the proposed method and the cost volume construction method of the correspondence response in [11] as an example. While the exact shapes of the cost profiles in the existing cost volume construction methods based on DFS are different from each other, the tendency that the cost profile has the minimum cost value at the corresponding disparity index is the same. Fig. 6(a) and (b) show the reference view image and the resulting disparity map of the proposed method, respectively. The cost profiles of dots (pixels) on the floor and the wall in Fig. 6(a) are shown in Fig. 6(c) (solid lines). A dot on the floor signifies the foreground, while a dot on the wall indicates the background. The low disparity index represents the foreground. Fig. 6(c) shows the profiles of the proposed method corresponding to the cost profile by [10] with the dotted line. In terms of the cost profiles, the proposed method can be considered as the slope sign of the cost profile. Vertical solid lines in Fig. 6(c) represent slices of $BW(x, y, d)$ with each fixed d as shown in Fig. 6(d). Each pixel has different accumulated result depending on its depth. The accumulated result at a pixel signifies the zero crossing index. The zero crossing index corresponds to the extremum index in the existing cost volume construction method. Given that the foreground is defined to have a positive sign, high value is assigned to the close scene on the resulting depth map of the proposed method. By assigning the positive value to the background, the depth ordering of the resulting depth map can be obtained similarly to the existing method.

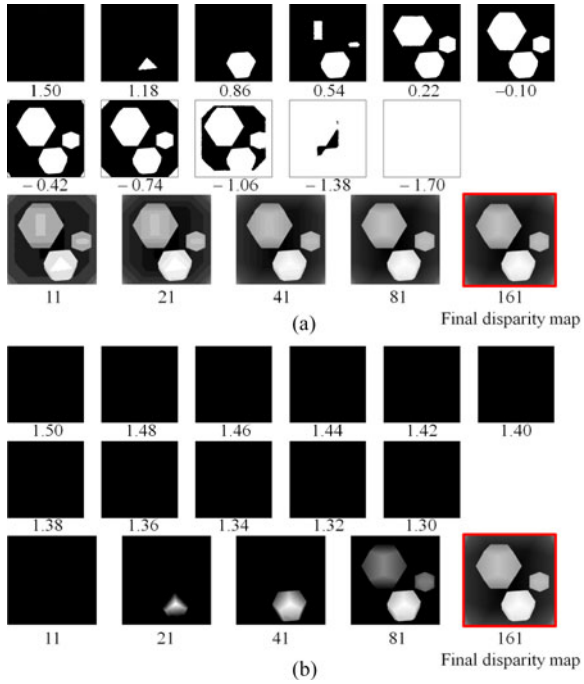


Fig. 7. Comparison of two different re-parameterization strategies. (a) Coarse-to-fine. (b) Front-to-back scanning.

From the framework of the proposed method (accumulation of binary maps), the proposed method has an advantage in terms of memory usage. Although Figs. 4 and 6 show the cost volume to explain the proposed method and to compare the proposed method with the existing methods, the proposed method actually does not need to construct the cost volume. By directly updating the computed binary map, the proposed method requires 2-D memory to obtain the disparity map.

D. Re-Parameterization Strategy in Cost Volume Construction

The cost volume is constructed by computing several cost slices with the re-parameterized light fields. In the local approach in DFS, Szeliski and Scharstein addressed the disparity space image (DSI) [30], which is a cost slice in the cost volume. They explained the computation of the cost slices as the sampling of the DSI. Similar to DFS, the computation of the cost slices in the light field can be considered as sampling of the DSI. That is, the light field re-parameterization on an arbitrary plane is to determine the sampling position. In this respect, various re-parameterization strategies can be used depending on the application because the disparity map can be obtained and displayed by accumulating the binary maps.

For example, Fig. 7 shows two different re-parameterization strategies using Platonic image in [28]. Fig. 7(a) and (b) show the coarse-to-fine and front-to-back re-parameterization strategies, respectively. The given disparity range of Platonic image (from –1.7 to 1.5) is quantized with a step size of 0.02 (total of 161 levels). The first and second rows are the 11 binary maps of each sampling strategy. Each binary map is obtained by

re-parameterizing the input light field with the values specified at the bottom of each image. The third rows are the disparity maps with the different numbers of binary maps that are specified at the bottom of each image. For a user, the coarse-to-fine re-parameterization strategy has an advantage to reduce the waiting time until the disparity map is displayed. The accumulation results of binary maps in the first and second rows in Fig. 7(a) and (b) are the first images in the third rows. Although the final disparity maps obtained by accumulating the 161 binary maps in Fig. 7(a) and (b) are the same, the intermediate results by accumulating 11, 21, 41, and 81 binary maps are different from each other. The coarse-to-fine re-parameterization strategy can effectively show the outline of the disparity map with only 11 binary maps while the front-to-back scanning re-parameterization strategy displays nothing with the first 11 binary maps.

Although two examples of the re-parameterization strategy are shown in Fig. 7, the proposed method can be used in various ways. It is also possible to re-parameterize the input light field in a back-to-front manner or in the disparity range of interest. If the surveillance system is focused on finding the object coming into a certain distance, the proposed method can be effectively used. That is, as the accumulation of 41 binary maps in Fig. 7(b), the disparity map of the certain distance range in the scene can be obtained. The existing methods can also compute the cost volume using the coarse-to-fine and front-to-back re-parameterization strategies, however, they cannot update and display disparity map during cost volume construction step.

IV. Experimental Results and Discussions

To evaluate the proposed method in various aspects, the synthetic and real images are used. The synthetic images in the dataset [28] provide a framework to compare the performance of the depth estimation methods. Using the stratified, test, and training images, the proposed method is compared with the six existing methods (EPI1 [15], LF [10], LF_OCC [12], SPO [31], RM3DE [32], and SC_GC [33]). Details of the resulting disparity maps of the proposed method (FBS) and the existing methods can be found in [34]. Using the real images from [10], [11], and [35], the qualitative evaluation of the proposed method and the existing methods is performed. In experiments, the proposed method is implemented in MATLAB on a PC with a 3.5GHz CPU and 32 GB of memory. In this environment, the average runtime for computing a single slice using the dataset [28] (size: $u \times v \times x \times y = 9 \times 9 \times 512 \times 512$) is 4.13 secs.

A. Performance Evaluation on Synthetic Images

Table I shows the averaged evaluation metrics with the synthetic images from the dataset [28]. Lower scores represent better performance for all metrics. The best performance for each metric is bold-faced. The proposed method shows reasonable performance compared to existing methods in terms of the mean squared error (MSE) and bad pixel (BadPix). Details of MSE and BadPix on each image are shown in Table II, in which the best performance for each image is bold-faced. In the proposed method, the radius and epsilon in the guided image filtering are 5 and 1e-5, respectively. In terms of the runtime, the performance of the proposed method is measured until the final disparity map is displayed. To obtain the disparity map using the proposed method, the given disparity range on each image is quantized with a step size of 0.02. For example, the disparity range [–1, 1] is quantized to a total of 101 levels. Although the performance of the proposed method is not outstanding in terms of the measured runtime in Table I, the proposed method can display outline of the disparity map during cost volume construction using coarse-to-fine re-parameterization strategy.

TABLE I
General and Photorealistic Performance Evaluation on Synthetic Images

	EPI1	LF	LF_ OCC	SPO	RM 3DE	SC_ GC	FBS
BadPix	24.32	16.19	17.58	**8.47**	10.21	14.30	13.29
MSE	5.98	9.13	6.70	3.97	3.92	6.58	**3.81**
Runtime (log10)	1.95	3.00	4.02	3.33	**1.67**	3.31	2.82
Bumpiness Cont. Surf.	2.45	1.32	2.04	2.04	1.88	**1.19**	2.12
Bumpiness Planes	2.07	1.68	1.87	2.30	1.75	**0.40**	2.27
Discont.	45.34	38.21	38.26	**22.81**	25.64	29.29	36.59
Fine Fattening	58.86	41.42	42.09	29.22	39.04	**25.19**	45.74
Fine Thinning	10.73	19.11	10.73	16.73	**10.28**	21.74	14.96

From the benchmark images, although the proposed method shows reasonable performance in terms of the general error metrics such as the MSE and BadPix, it shows relatively poor performance than the existing methods in the photorealistic aspect. The existing methods have a step to optimize the cost volumes or initial depth maps for improving the performance by using various constraints, such as smoothness and edge. However, the proposed method does not have the optimization step (global method in stereo). It can be viewed as a step of cost volume construction in the existing methods and the accumulated final disparity map can be viewed as the initial disparity map in the existing methods [6]. It might be fairer to compare the proposed method with the initial disparity map in the existing methods than with the final disparity map. The proposed method without an optimization step shows competitive results in terms of the general error metrics compared to the existing methods with the optimization step. Specifically, it has more room for improvement in view of photorealistic performance evaluation if the optimization step is additionally used.

Fig. 8 shows the resulting disparity maps of the stratified images. Details of the resulting disparity maps of the proposed method and the existing methods can be found in [34]. Fig. 9 shows xd slices of the proposed method on Dots image for further analysis. In Dots image, there are 15 identical grid cells with Gaussian noise. The variances of Gaussian noise are linearly growing between 0.0 and 0.2 in row-major order [28]. Fig. 9(a) shows the reference view image. Fig. 9(b) shows xd slices of C_V along the lines in Fig. 9(a). Fig. 9(c) shows xd slices of $\tilde{C}_V$ and BW with two different radius parameters of the guided image filter. The first three slices at the top in Fig. 9(b) and (c)

TABLE II
PERFORMANCE EVALUATION ON EACH SYNTHETIC IMAGE

Category	Image	MSE							BadPix						
		EPI1	LF	LF_ OCC	SPO	RM 3DE	SC_ GC	FBS	EPI1	LF	LF_ OCC	SPO	RM 3DE	SC_ GC	FBS
Stratified	Backgammon	9.56	13.01	21.59	**4.59**	9.21	5.65	5.67	21.33	5.52	19.01	**3.78**	7.51	4.89	10.31
	Dots	5.73	5.68	3.30	5.24	3.29	19.81	**2.09**	62.00	**2.90**	5.82	16.27	31.74	48.51	12.84
	Pyramids	**0.027**	0.273	0.098	0.04	0.06	0.07	0.029	0.858	12.35	3.172	0.86	0.68	1.23	**0.55**
	Stripes	2.67	17.45	8.13	6.96	**1.00**	15.62	1.32	25.81	35.74	18.41	14.99	**4.13**	36.71	19.12
Training	Boxes	8.72	17.43	9.85	9.11	**7.63**	12.29	8.90	24.45	23.02	26.52	**15.89**	19.26	18.57	27.15
	Cotton	2.25	9.17	1.07	1.31	**0.34**	1.87	0.76	13.93	7.83	6.22	2.59	**2.02**	6.28	5.01
	Dino	1.23	1.16	1.14	**0.31**	0.36	0.50	0.66	10.35	19.03	14.91	**2.18**	3.89	6.03	7.55
	Sideboard	2.85	5.07	2.30	**1.02**	1.07	1.82	1.16	18.38	21.99	18.50	9.30	**8.48**	9.74	13.92
Test	Bedroom	0.57	0.47	0.63	0.21	**0.20**	0.26	0.48	13.59	13.86	17.57	**4.86**	5.10	5.06	5.77
	Bicycle	8.52	11.73	7.67	**5.57**	6.54	5.88	5.81	25.21	19.79	21.56	**10.91**	16.08	12.21	19.66
	Herbs	24.70	21.34	22.20	**11.23**	15.52	13.02	15.46	47.08	18.11	36.83	**8.26**	10.14	12.69	14.75
	Origami	5.01	6.76	2.30	**2.03**	1.84	2.23	3.32	28.90	14.18	22.43	11.70	13.56	**10.67**	22.86
Average		5.98	9.13	6.69	3.97	3.92	6.58	**3.81**	21.33	16.19	17.58	**8.47**	10.22	14.30	13.29

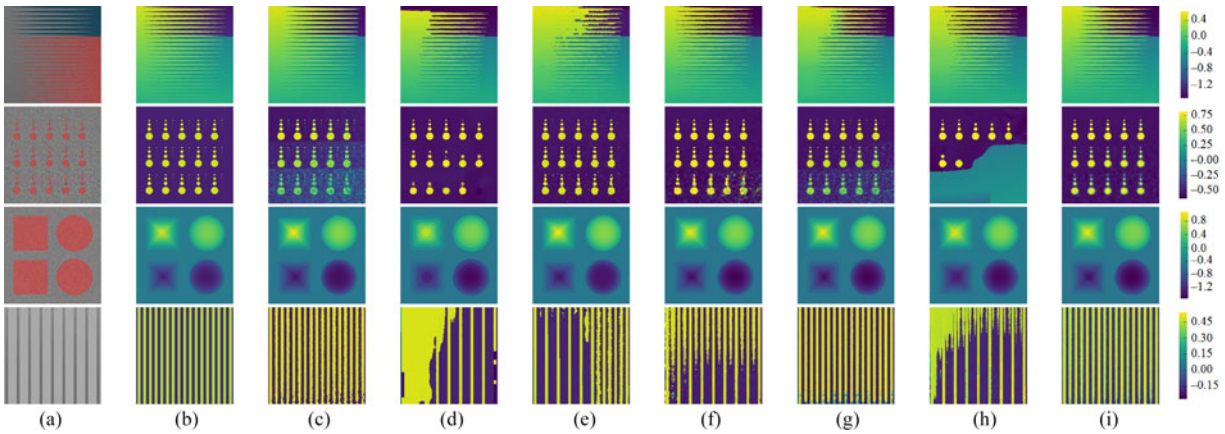


Fig. 8. Comparison of the resulting disparity maps on the stratified images. (a) Reference view image. (b) Ground truth. (c) EPI1 [15]. (d) LF [10]. (e) LF_OCC [12]. (f) SPO [31]. (g) RM3DE [32]. (h) SC_GC [33]. (i) FBS.

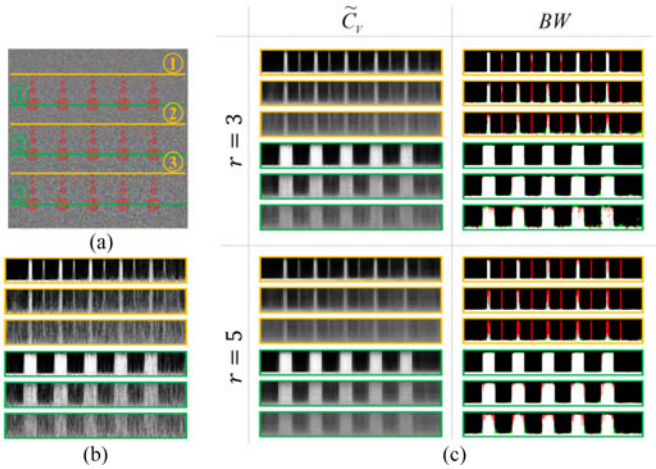


Fig. 9. xd slices of the proposed method on Dots image from [28]. (a) Reference view. (b) xd slices of $C_V(x, y, d)$. (c) xd slices of $\tilde{C}_V(x, y, d)$ and $BW(x, y, d)$ with different radius parameters in the guided image filter.

correspond to the right-numbered lines in Fig. 9(a), whereas the other three slices at the bottom correspond to the left-numbered lines. Each slice of BW is overlapped with the corresponding ground truth generated using the provided ground truth disparity map. The incorrectly separated foreground and background are represented by the green and red pixels, respectively. As shown in Fig. 9(b), the slices of C_V are more degraded if the noise level increases. As shown in Fig. 9(c), the higher the noise level is, the more unclear boundaries between foreground and background in the slices $\tilde{C}_V$ and BW are. While BW using the guided image filter with smaller radius is more degraded by the noise than that with larger radius, BW using the guided image filter with smaller radius shows better performance in terms of the missed dot.

B. Performance Evaluation on Real Images

Fig. 10 shows the performance comparison of the proposed method with the existing methods on real images. In the proposed method, the disparity range for each image is empirically determined and uniformly quantized to 20 levels. The images in Fig. 10(a) are from dataset [35]. The resulting depth maps of the proposed method are shown in the second column, and those of the existing methods [10], [12], and [36] are shown in the

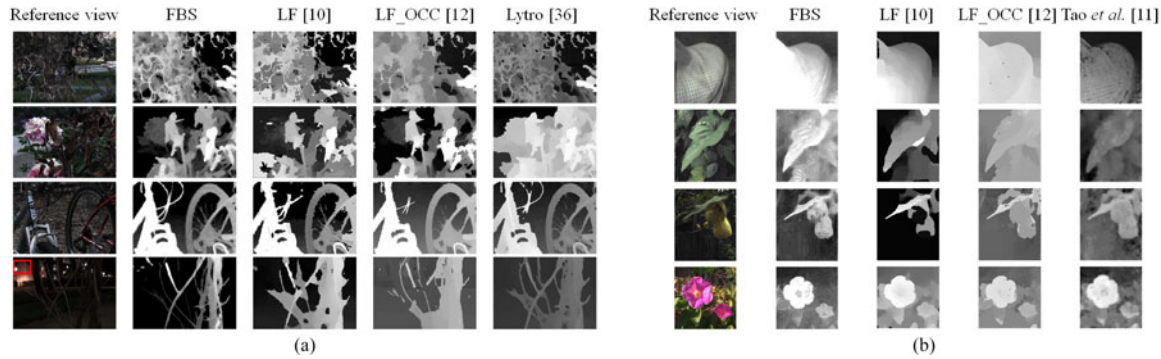


Fig. 10. Comparison of the proposed method and the existing methods on real images. (a) Dataset [35]. (b) Datasets [10, 11].

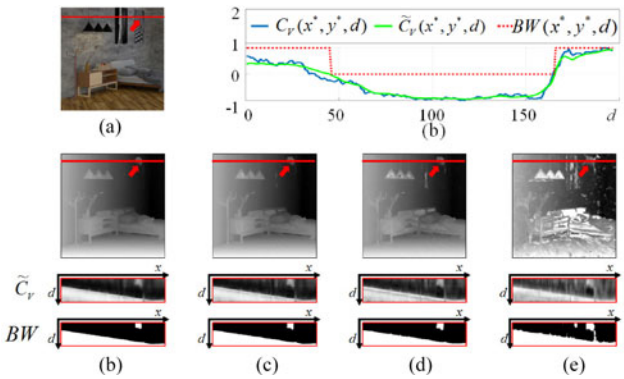


Fig. 11. Limitation of the proposed method (sign reversal). (a) Reference view of Bedroom image. (b) A profile of a pixel on the region indicated by the arrow. (c)–(f) Intermediate results and accumulated disparity maps using the light field with 9×9,7×7,5×5, and 3×3 angular resolutions, respectively.

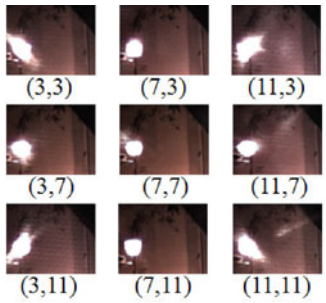


Fig. 12. Influence of the non-Lambertian object on the proposed method.

third, fourth, and fifth columns, respectively. Fig. 10(b) shows the experimental results on the images provided by [10] and [11]. The reference images are shown in the first column. The second column shows the resulting depth maps of the proposed method. Those of the existing methods [10]–[12] are shown in the third, fourth, and fifth columns, respectively.

For each image, the proposed method can use the dynamic range more effectively than the existing methods. If thin structures are larger than the angular resolution, the proposed method shows better performance in terms of preservation of the occlusion boundary and detail structure of the input image. However, as shown in the third depth map in Fig. 10(a), thin structures, such as the wheel spokes of a bicycle, which are smaller than the angular resolution, are not well-preserved as discussed in the case of the synthetic images in Fig. 9.

C. Limitation of the Proposed Method

As the initial disparity map, the proposed method does not have many problems in contrast to the initial disparity maps in the existing methods. However, the proposed method also has some problems on two cases.

First, the proposed method incorrectly estimates the disparity on the region with the repeated pattern if the input light field is re-parameterized with a relatively large disparity value due to the angular aliasing. The influence of the angular aliasing is increased if the angular resolution is low. Fig. 11(a) shows the reference view image. Fig. 11(b) shows profiles of a light field with 9×9 angular resolution at a pixel in the region that the arrow indicates. The super-scripts $*$ in the legend indicate that the corresponding arguments are fixed. Fig. 11(c)–(f) show the resulting disparity maps and the intermediate results using the light field with 9×9, 7×7, 5×5, and 3×3 angular resolutions, respectively. As shown in Fig. 11(c), the stained artifacts are produced on the resulting disparity map indicated by the arrow even though the full angular resolution is used. In Fig. 11(b), although the true disparity index of this region is near 170, the sign of the disparity index smaller than 50 is reversed, which causes the artifacts in the resulting disparity map. In Fig. 11(c)–(f), xd slices of $\tilde{C}_V(x, y, d)$ and $BW(x, y, d)$ along the line in the resulting disparity maps are shown at the bottom of each resulting disparity map. The lower angular resolution is used, the larger region is influenced by the angular aliasing. Especially, aliasing artifacts are significantly increased if the light field with 3×3 angular resolution is used.

Second, the proposed method incorrectly estimates the disparity map on the non-Lambertian region as shown in the last row of Fig. 10(a). The red-boxed region is enlarged in Fig. 12 with different views, where light streaks are differently observed in sub-aperture images and the opposite disparity signs are produced in those regions. Thus, the foreground and background are incorrectly separated and the artifacts are introduced on the resulting disparity map as shown in the last row of Fig. 10(a).

V. Conclusion

We propose a depth estimation method from a light field by accumulating binary maps based on the foreground- background separation. An optical phenomenon that the bundles of

rays from the background are flipped is used in the proposed method. To explain how the disparity map can be obtained by simple accumulation, the relations to the existing cost volume construction methods are shown. Experimental results on the synthetic and real images show that the proposed method reasonably performs in terms of general quantitative evaluation. The proposed method has advantages in terms of memory usage and re-parameterization strategies. Future work will focus on improving the photorealistic performance of the proposed depth estimation method.

REFERENCES

[1] J. Wan, G. Guo, and S. Z. Li, "Explore efficient local features from RGB-D data for one-shot learning gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1626–1639, Aug. 2016.

[2] G. Seguin, K. Alahari, J. Sivic, and I. Laptev, "Pose estimation and segmentation of multiple people in stereoscopic movies," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1643–1655, Aug. 2015.

[3] C. Zhou and S. K. Nayar, "Computational cameras: Convergence of optics and processing," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3322–3340, Dec. 2011.

[4] M. Levoy, "Light fields and computational imaging," *IEEE Comput.*, vol. 39, no. 8, pp. 46–55, Aug. 2006.

[5] R. Ng, M. Levoy, M. Bredif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep. CSTR 2005-02, Apr. 2005.

[6] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.

[7] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2806–2813.

[8] T. Ryu, B. Lee, and S. Lee, "Mutual constraint using partial occlusion artifact removal for computational integral imaging reconstruction," *Appl. Opt.*, vol. 54, no. 13, pp. 4147–4153, May 2015.

[9] C. Chen, H. Lin, Z. Yu, S. B. Kang, and J. Yu, "Light field stereo matching using bilateral statistics of surface cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1518–1525.

[10] H.-G. Jeon *et al.*, "Accurate depth map estimation from a lenslet light field camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 1547–1555.

[11] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 673–680.

[12] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Depth estimation with occlusion modeling using light-field cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2170–2181, Nov. 2016.

[13] W. Williem and I. K. Park, "Robust light field depth estimation for noisy scene with occlusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 4396–4404.

[14] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4D light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 41–48.

[15] O. Johannsen, A. Sulc, and B. Goldluecke, "What sparse light field coding reveals about scene structure," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 3262–3270.

[16] D. Dansereau and L. Bruton, "Gradient-based depth estimation from 4-D light fields," in *Proc. IEEE Int. Symp. Circuits Syst.*, Vancouver, BC, Canada, May 2004, pp. 549–552.

[17] A. Isaksen, L. McMillan, and S. J. Gortler, "Dynamically reparameterized light fields," in *Proc. SIGGRAPH Int. Conf. Comput. Graph. Interact. Techn.*, New Orleans, LA, USA, Jul. 2000, pp. 297–306.

[18] J. Y. Lee and R.-H. Park, "Separation of foreground and background from light field using gradient information," *Appl. Opt.*, vol. 56, no. 4, pp. 1069–1078, Feb. 2017.

[19] T. E. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 972–986, May 2012.

[20] A. Sellent and P. Favaro, "Which side of the focal plane are you on?," in *Proc. IEEE Int. Conf. Comput. Photography*, Santa Clara, CA, USA, May 2014, pp. 1–8.

[21] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1, pp. 7–42, Apr. 2002.

[22] O. Johannsen *et al.*, "A taxonomy and evaluation of dense light field depth estimation algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Honolulu, HI, USA, Jul. 2017, pp. 82–99.

[23] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. SIGGRAPH Int. Conf. Comput. Graph. Interact. Techn.*, New Orleans, LA, USA, Aug. 1996, pp. 31–42.

[24] A. Levin and F. Durand, "Linear view synthesis using a dimensionality gap light field prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 1831–1838.

[25] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[26] T. Ojala, M. Peitkainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[27] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[28] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. Asian Conf. Comput. Vis.*, Taipei, Taiwan, Nov. 2016, pp. 19–34.

[29] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," in *Proc. Vis., Model. Vis.*, Lugano, Switzerland, Sep. 2013, pp. 225–232.

[30] R. Szeliski and D. Scharstein, "Sampling the disparity space image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 3, pp. 419–425, Mar. 2004.

[31] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Comput. Vis. Image Understanding*, vol. 145, pp. 148–159, Apr. 2016.

[32] A. Neri, M. Carl, and F. Battisti, "A multi-resolution approach to depth field estimation in dense image arrays," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 3358–3363.

[33] L. Si and Q. Wang, "Dense depth-map estimation and geometry inference from light fields via global optimization," in *Proc. Asian Conf. Comput. Vis.*, Taipei, Taiwan, Nov. 2016, pp. 83–98.

[34] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "4-D light field benchmark." 2017. [Online]. Available: http://lightfield-analysis.net

[35] A. S. Raj, M. Lowney, R. Shah, and G. Wetzstein, "Stanford Lytro light field archive," 2016. [Online]. Available: http://lightfields.stanford.edu

[36] Lytro, "The Lytro camera." 2017. [Online]. Available: http://lytro.com

Jae Young Lee received the B.S. and M.S. degrees from the Department of Electronic Engineering, Sogang University, Seoul, South Korea, in 2015 and 2017, respectively. He is currently affiliated with LG Innotek, Seoul. His current research interests include image processing and computer vision.

Rae-Hong Park (S'76–M'84–SM'99) received the B.S. and M.S. degrees in electronics engineering from Seoul National University, Seoul, South Korea, in 1976 and 1979, respectively, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 1981 and 1984, respectively.

In 1984, he joined the faculty of the Department of Electronic Engineering, Sogang University, Seoul, where he is currently a Professor. In 1990, he spent his sabbatical year as a Visiting Associate Professor with the Computer Vision Laboratory, Center for Automation Research, University of Maryland at College Park. In 2001 and 2004, he spent sabbatical semesters at Digital Media Research and Development Center (DTV image/video enhancement), Samsung Electronics Co., Ltd., Suwon, South Korea. In 2012, he spent a sabbatical year in Digital Imaging Business (R&D Team) and Visual Display Business (R&D Office), Samsung Electronics Co., Ltd., Suwon. His current research interests include computer vision, pattern recognition, and video communication.

He served as the Editor for the Korea Institute of Telematics and Electronics (KITE) *Journal of Electronics Engineering* from 1995 to 1996. He received the 1990 Postdoctoral Fellowship presented by the Korea Science and Engineering Foundation, the 1987 Academic Award presented by the KITE, the 2000 Haedong Paper Award presented by the Institute of Electronics Engineers of Korea, the 1997 First Sogang Academic Award, and the 1999 Professor Achievement Excellence Award presented by Sogang University. He is a co-recipient of the Best Student Paper Award of the IEEE International Symposium on Multimedia (ISM 2006) and IEEE International Symposium on Consumer Electronics (ISCE 2011).

Occlusion-Model Guided Antiocclusion Depth Estimation in Light Field

Hao Zhu, *Student Member, IEEE*, Qing Wang, *Member, IEEE*, and Jingyi Yu

***Abstract*—Occlusion is one of the most challenging problems in depth estimation. Previous work has modeled the single-occluder occlusion in light field and achieves good performances, however it is still difficult to obtain accurate depth for multioccluder occlusion. In this paper, we explore the complete occlusion model in light field and derive the occluder-consistency between the spatial and angular spaces, which is used as a guidance to select unoccluded views for each candidate occlusion point. Then, an antiocclusion energy function is built to regularize the depth map. Experimental results on both synthetic and real light-field datasets have demonstrated the advantages of the proposed algorithm compared with state-of-the-art algorithms of light-field depth estimation, especially in multioccluder cases.**

***Index Terms*—Anti-occlusion depth estimation, light field, occlusion model.**

I. Introduction

DEPTH estimation is one of the most important problems in computer vision and has been studied for decades. It has a wide range of applications [1] in tasks of saliency detection [2], [3], autonomous vehicle [4], seeing through occlusion [5], [6], light field segmentation [7], [8] and novel view synthesis [9], [10]. Most algorithms operate under a common assumption that the color of a point observed from different views ought to be the same or similar, i.e., Lambertian reflectance. This assumption holds for diffuse surfaces. However, it fails for occlusion points due to the under-sampling in the angular space [11] which leads that these points cannot be observed from all views. The estimated depths of occlusion points under the photo-consistency assumption tend to be over-smoothed (Fig. 1), which will decrease the accuracy for other tasks.

In literatures, there are a great number of approaches to handle occlusions. Kolmogorov *et al.* [12] encoded the visibility constraint and introduced an occlusion term to smooth it. Woodford *et al.* [13] added an additional second order smoothness term and used Quadratic Pseudo-Boolean Optimization to solve it. Then, Bleyer *et al.* [14] applied the asymmetric occlusion model to improve depth estimation. However, due to the wide baseline and the lack of views, these works only realized the visibility and invisibility of occlusion boundaries but omitted the distribution of un-occluded views. As a result, heavy occlusion cannot be well handled.

Light field camera [15]–[18] captures hundreds of views in a single shot, which provides new sights for occlusion-aware depth estimation. Wanner *et al.* [19] applied structure tensor to analyze the Epipolar Plane Image (EPI). This method only took advantage of angular samples in one dimension. The tensor field becomes too random to analyze in heavy occlusion. Yu *et al.* [20] encoded the constraints of 3D lines and introduced line assisted graph cuts to improve depth estimation. However, the 3D lines are partitioned into small and incoherent segments in heavy occlusion, which leads to bad estimation. Chen *et al.* [21] proposed to select un-occluded views by using a bilateral metric in the angular space. However, this selection lacks the guidance of physical model and the number of un-occluded views is also need to preset. In addition, Tao *et al.* [22] modeled the defocus and correspondence cues in light field and combined them into a Markov Random Field to optimize depth map. Jeon *et al.* [23] reduced the astigmatism [24] in light field camera using a phase-shift based sub-pixel displacement and improved depth accuracy by introducing feature correspondence into global optimization. However, both two algorithms do not consider the occlusion and always provide over-smooth results in occlusion boundaries.

More recently, Wang *et al.* [25] analyzed the formation of occlusion and found the projection consistency between the spatial and angular patches in occlusion boundaries. A single-occluder occlusion in light field is proposed and it is utilized to separate the occluded/un-occluded views in a simple way, i.e., half and half using a straight line. However, the method has two disadvantages. First, the size of spatial patch is arbitrary and it is always predefined. Due to the uncertainty of patch size, the second issue appears. Different patch sizes may lead to different occlusion types. A multi-occluder occlusion becomes a single-occluder occlusion with a small patch size and vice versa. The single-occluder occlusion assumption is obviously unsuitable for multi-occluder occlusion. As shown in Fig. 1, although the thin structure of wires is reconstructed well using [25], the intersections of wires are sometimes missing or over-smoothed. The main reason is that, the intersection of wires is a multi-occluder occlusion and the single-occluder occlusion assumption does not hold in this case.

Manuscript received December 30, 2016; revised April 23, 2017; accepted July 3, 2017. Date of publication July 21, 2017; date of current version October 23, 2017. The work was supported in part by National Natural Science Foundation of China under Grants 61531014, 61401359, and 61272287. The guest editor coordinating the review of this paper and approving it for publication was Prof. Diego Gutierrez. *(Corresponding author: Hao Zhu.)*

H. Zhu and Q. Wang are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: zh920307@outlook.com; qwang@nwpu.edu.cn).

J. Yu is with the ShanghaiTech University, Shanghai 200031, China (e-mail: jingyi.udel@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2730818

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

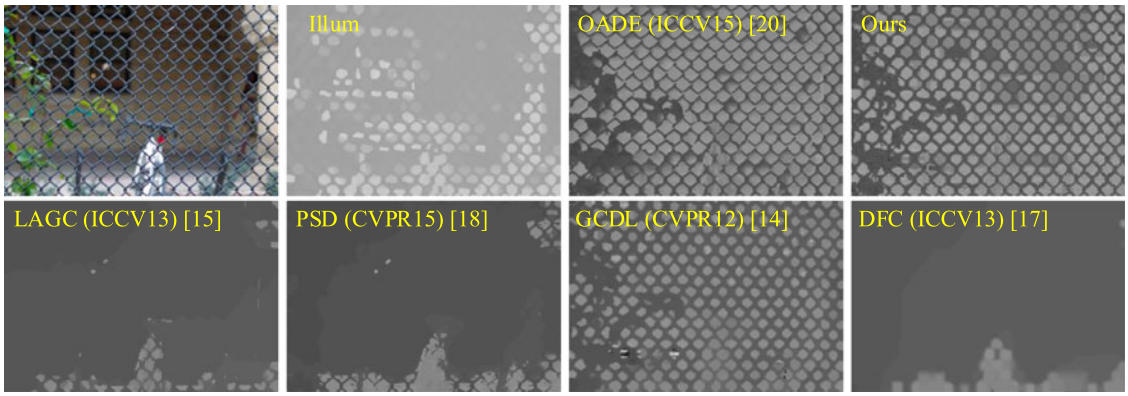


Fig. 1. Comparisons of depth estimation of different algorithms on real light field data. Only the methods of [25], [19] and ours can recover exact shape of wire mesh. However, the wire is over-smoothed by [19] and the intersections of wires are missing by [25]. Our method retains both the thin structure and intersections.

In this paper, we explore the complete light field occlusion model which is suitable for both single-occluder and multi-occluder occlusions, and derive the obtention of spatial patch size in projection consistency. The main contributions are as follows,

1) The complete light field occlusion model.
2) An algorithm to accurately select the un-occluded views in the angular space with the guidance of occlusion model which does not need any priors.
3) A depth estimation algorithm which can robustly preserve occlusion boundaries.

In Section II, we firstly model the multi-occluder occlusion in light field and derive the occlusion-consistency between the spatial and angular spaces, i.e., the occluded views in the angular space are projections of the occluder in the spatial space. Then the patch size of projection in the spatial space is derived using the refocus theory. With the guidance of occluder-consistency, we select un-occluded views for each candidate occlusion point using a partition in the spatial patch in Section III-A, and obtain an initial depth map using un-occluded views in Section III-B. Then, we detect occlusion points from initial depth map according to visible constraint in Section IV-A, and improve the accuracy of un-occluded views selection in Section IV-B. Finally the occlusion map and un-occluded views are used to build an anti-occlusion energy function to refine depth map in Section IV-C. In Section V, we demonstrate complete comparisons with state-of-the-art algorithms quantitatively and qualitatively, showing great advantages compared with previous works [19]–[23], [25].

II. The Light Field Occlusion Model

In this section, we first analyze the formation of the occlusion based on the physical model and then explore the complete light field occlusion model, including single-occluder and multi-occluder occlusions.

A. Definitions and Notations

Before building light field occlusion model, we first give some definitions and notations of light field model.

TABLE I
Notations of Light Field Occlusion Model

Term	Definition
(u, v)	angular/camera plane coordinates
(x, y)	spatial/image plane coordinates
(X, Y, Z)	world coordinates
$\vec{e}_A$	a vector in the angular coordinate system
$\vec{e}_I$	a vector in the spatial coordinate system
$\vec{e}_W$	a vector in the world coordinate system

Definition 1: Single-occluder occlusion refers to the occlusion that there is only one object occluding the background and the occlusion boundary is approximately a straight line.

Definition 2: Multi-occluder occlusion refers to the occlusion that there are more than one object occluding the background or one object with an irregular occlusion boundary.

A point (x, y, u, v) in the 4D light field describes a ray emitting from the world point (X, Y, Z) and intersecting with two parallel planes, namely, the uv camera plane and the xy image plane, which we refer to as the angular and spatial spaces respectively. Table I presents a list of terms used throughout this section.

B. Occlusion Model

Previous work [25] has proved the occluder-consistency for single-occluder occlusion, i.e., when refocused to the correct depth, the edge which separates the occluded and un-occluded views in the angular patch has the same orientation as the occlusion edge in the spatial patch. This property is useful for single-occluder occlusions, however it fails in multi-occluder situation as the occluded and un-occluded pixels in the angular patch cannot be simply divided into two regions by a straight line.

Now we first analyze a simple multi-occluder occlusion shown in Fig. 2. Considering a point (X_0, Y_0, F) on the focal plane (the left part of Fig. 2(a)), an occluder intersects at (X_0, Y_0, Z_0) $(0 < Z_0 < F)$. Note that the occluder has two edges and directional vectors of these two edges on the plane

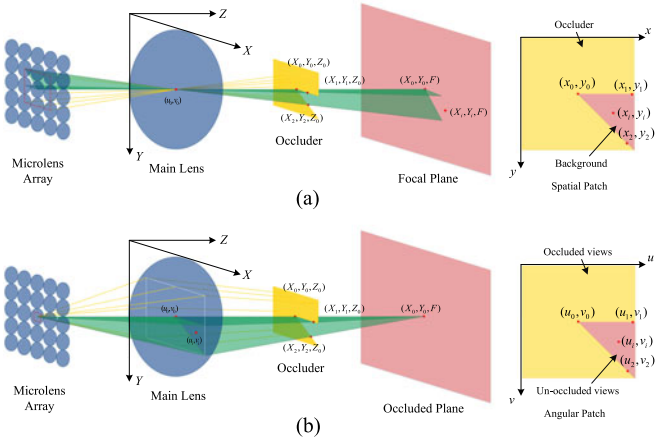


Fig. 2. The light field occlusion model. The left part of (a) denotes pinhole imaging process at the central view (u_0, v_0). The right-most of (a) shows a spatial patch centered at (x_0, y_0) under this view. The left part of (b) denotes a refocusing at the occluded plane with the depth of F. Only the views within dihedral angle (by two green planes) can observe the 3D point (X_0, Y_0, F) whilst other views are blocked by the occluder. The right-most of (b) shows the corresponding angular patch of (x_0, y_0). (a) Pinhole model. (b) "Reversed" pinhole model.

$Z = Z_0$ are,

$$\begin{aligned}\vec{e}_W^1 &= (k_{X_1}, k_{Y_1}) = (X_1 - X_0, Y_1 - Y_0)\\ \vec{e}_W^2 &= (k_{X_2}, k_{Y_2}) = (X_2 - X_0, Y_2 - Y_0).\end{aligned} \tag{1}$$

The larger angle between these two vectors denotes occluded areas (the yellow areas in Fig. 2). Without loss of generality, we assume $k_{Y_j} > 0, j = 1, 2$. For any point (X_i, Y_i, F) on the focal plane, it is observed by the view (u_0, v_0) iff it meets the following inequalities,

$$\begin{aligned}k_{Y_1}(X_i - X_0) - k_{X_1}(Y_i - Y_0) > 0\\ k_{Y_2}(X_i - X_0) - k_{X_2}(Y_i - Y_0) < 0.\end{aligned} \tag{2}$$

We then project these inequalities from the world coordinate system to image system (the right-most image of Fig. 2(a)). The corresponding vectors of $\vec{e}_W^1$ and $\vec{e}_W^2$ are $\vec{e}_I^1 = (k_{x_1}, k_{y_1})$ and $\vec{e}_I^2 = (k_{x_2}, k_{y_2})$ respectively. Obviously, we have $\vec{e}_I^1 = \lambda_1 \vec{e}_W^1$ and $\vec{e}_I^2 = \lambda_1 \vec{e}_W^2$, where λ_1 is a scale factor to control the scaling relationship between the world and image coordinate systems. For any pixel (x_i, y_i) on the image plane, it is a background pixel iff

$$\begin{aligned}k_{y_1}(x_i - x_0) - k_{x_1}(y_i - y_0) > 0\\ k_{y_2}(x_i - x_0) - k_{x_2}(y_i - y_0) < 0.\end{aligned} \tag{3}$$

Then considering the main lens plane (the left part of Fig. 2(b). The light field is refocused to the depth of F). For any view (u_i, v_i) on the main lens plane, it can acquire the point (X_0, Y_0, F) iff

$$\begin{aligned}k_{v_1}(u_i - u_0) - k_{u_1}(v_i - v_0) > 0\\ k_{v_2}(u_i - u_0) - k_{u_2}(v_i - v_0) < 0,\end{aligned} \tag{4}$$

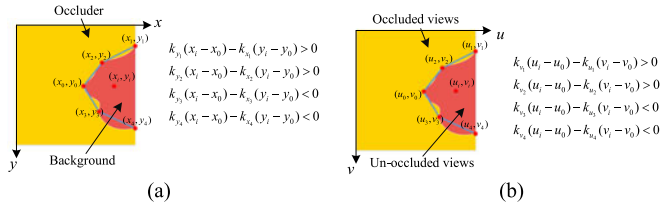


Fig. 3. Complex occlusion boundary. The yellow and red areas correspond to (a) occluder and background, (b) occluded and un-occluded views, respectively. The boundaries of occluder and occluded views are irregular. Four blue line segments are used to represent the irregular boundary. For each pixel-view pair $((x, y), (u, v))$, it is a background pixel/un-occluded view iff it meets four inequalities. These two groups of inequalities meet the one-to-one correspondence constraint, thus Prop.1 holds. When more line segments are used to fit the boundary for better accuracy, the inequalities still meet the one-to-one correspondence constraint, and so does Prop.1. (a) Spatial patch. (b) Angular patch.

where $\vec{e}_A^j = (k_{u_j}, k_{v_j})$ and $\vec{e}_A^j = \lambda_2 \vec{e}_I^j$, $j = 1, 2$. λ_2 is a scale factor to control the scaling relationship between the image and angular coordinate systems.

Revisiting (3) and (4), it can be seen that they have same inequalities with one-to-one correspondence since (u_0, v_0) and (x_0, y_0) are on the same line. Based on the above-mentioned facts, the following proposition is obtained accordingly.

Proposition 1: The occluded views in the angular space are projections of the occluder in the spatial space.

In other words, in a spatial patch, the corresponding views of the occluder are the occluded views, and the corresponding views of the background are the un-occluded views. This proposition is called the occluder-consistency later.

Note that the proposition mentioned above is obtained under a simple multi-occluder assumption. For a more complex multi-occluder, its irregular occlusion boundary can be approximately represented by a set of line segments (see Fig. 3). Correspondingly, (3) and (4) will contain more inequalities. The number of inequalities is equal to the number of these line segments.

In a short summary, it is noticed that the single-occluder occlusion model in [25] is just a special case of our model, i.e., the occlusion boundary is approximately modeled using a straight line. The proposed occlusion model is a complete model in light field.

C. Projection Radius

For the occluded points, the occluded views in the angular space are projections of the occluder in the spatial space in a local patch. For simplicity, we derive the radius of the patch in a 2D light field. In Fig. 4, the purple thick line denotes the background at depth α_0 and the orange thick line denotes the occluder at depth α_1. The blue line denotes the camera plane and x is a point in the background. $u_k, k = 0, \ldots, 4$ denote different views in light field.

Firstly, the light field is refocused to the background at depth α_0 [26] (Fig. 4(a))[1],

$$LF_{\alpha_0}(x, u) = LF_0\left(x + (u - u_2)\left(1 - \frac{1}{\alpha_0}\right), u\right), \tag{5}$$

[1]The coefficient α of x is often ignored in computer vision tasks, see [22], [25].

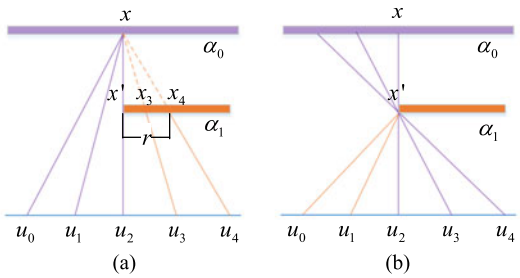


Fig. 4. The derivation of projection radius between the spatial and angular patches. In (a), the light field is refocused to the background at depth α_0, where the light rays from u_0, u_1, u_2 converge to point x and the light rays from u_3, u_4 are blocked by the occluder. In (b), the light field is refocused to the foreground at depth α_1. Light rays from all views converge to the point x'.

where LF_0 is the input light field, LF_{α_0} is the refocused light field at depth α_0, and u_2 is the central view of light field. It is noticed that light rays from views u_0, u_1, u_2 converge to the point x. Meanwhile, light rays from u_3, u_4 are blocked by the occluder. Correspondingly, u_3, u_4 are the occluded views in the angular space and the images of these two views come from points x_3, x_4. In other words, the horizontal distance r between x_4 and x is the projection radius.

Then, the light field is refocused to the foreground at depth α_1 (Fig. 4(b))

$$LF_{\alpha_1}(x,u) = LF_0\left(x + (u-u_2)\left(1-\frac{1}{\alpha_1}\right), u\right), \quad (6)$$

where LF_{α_1} is the refocused light field at depth α_1. It can be seen that light rays from all views converge to the point x'. As the horizontal distance between x' and x is 0, the projection radius r is obtained

$$\begin{aligned} r &= \left| x + (u_4-u_2)\left(1-\frac{1}{\alpha_1}\right) - x - (u_4-u_2)\left(1-\frac{1}{\alpha_0}\right)\right| \\ &= \left|(u_4-u_2)\left(\frac{1}{\alpha_0}-\frac{1}{\alpha_1}\right)\right|. \end{aligned} \quad (7)$$

The projection radius is a product of the angular radius of light field and the difference of disparities between the background and occluder.

III. Depth Initialization

In this section, we first present how to use the occluder-consistency between the spatial and angular spaces to select the un-occluded views and then introduce how to obtain an initial depth map using these un-occluded views.

A. Un-Occluded Views Selection

We first give an important assumption about occlusion in the proposed algorithm. The occluder has a different color of the occluded point. For the situation that the occluder is similar to occluded point, as far as we know, no work can handle it. Based on this assumption, the following proposition holds,

Proposition 2: An occlusion point is surely an edge point but an edge point may not be an occlusion point.

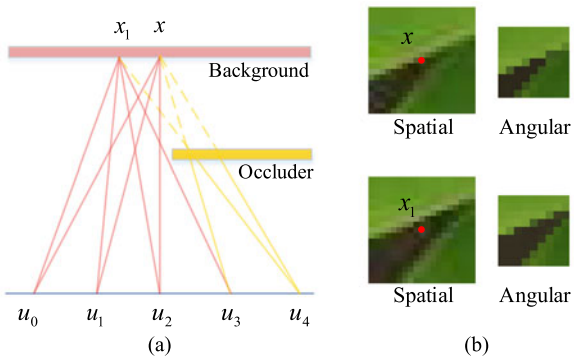


Fig. 5. The occlusion in different views. (a) demonstrates the model of occlusion in the central view and other views respectively. Point x is occluded in the central view and point x_1 is occluded in other views. (b) demonstrates the image of occlusion in the central view and other views respectively. It is noticed that x is an edge point and there are many occluded views in its angular patch when refocused to its true depth. The point x_1 is not an edge point, however as it is near the edges, there are many occluded views in its angular patch too.

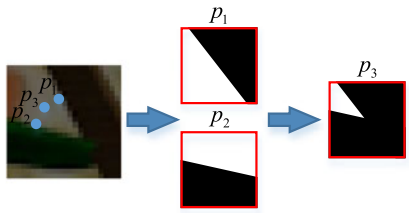


Fig. 6. The voting strategy of obtaining un-occluded views for pixels occluded in other views. The first col shows that points p_1, p_2 are occluded in central view and p_3 is occluded in other views. The second row shows the un-occluded views of p_1, p_2 (the areas labeled with white are un-occluded views). The third col is the un-occluded views of p_3 after voting.

With Prop. 2, the Canny edge detector is firstly applied to find the candidate occlusion points S_{occ}. Then the K-means clustering [27] is applied for the local image patch[2] centered at each occlusion point p from S_{occ} (the feature is the RGB color, and the number of labels is 2). For each patch, the pixels that share the same label with central pixel are labeled as background or un-occluded points. According to the occluder-consistency mentioned in Prop. 1, corresponding views in the angular patch of central pixel p are labeled as un-occluded views Ω_p^{uocc}. These pixels are candidate occlusion pixels in the central view (Fig. 5). For pixels occluded in other views, the un-occluded views are voted from its neighborhood system (Fig. 6).

B. Depth Estimation

With the un-occluded views selection, a robust initial depth map is obtained based on classical photo-consistency in the un-occluded views.

We refocus the light field to different depth α,

$$LF_\alpha(x,y,u,v) = LF_0(x + u\left(1-\frac{1}{\alpha}\right), y + v\left(1-\frac{1}{\alpha}\right), u, v), \quad (8)$$

[2]The patch size is set as half of the angular resolution of light field initially since we do not have disparity map here.

where LF_0 is the input 4D light field and LF_α is the refocused light field in depth α. (x, y) is the spatial coordinate while (u, v) is the angular coordinate. Then, the matching cost of each pixel p is defined as,

$$C_\alpha^{uocc}(p) = \frac{1}{|\Omega_p^{uocc}|} \sum_{(u,v)\in\Omega_p^{uocc}} |A_{p_\alpha}(u,v) - A_{p_\alpha}(0,0)|, \quad (9)$$

where Ω_p^{uocc} is the set of un-occluded views (Section III-A) and $|\cdot|$ denotes its size. $A_{p_\alpha}(u,v)$ is angular image of pixel p at depth α, i.e., $A_{p_\alpha}(u,v) = LF_\alpha(x_p, y_p, u, v)$. $A_{p_\alpha}(0,0)$ is the intensity of pixel p in the central view.

Then, the initial depth of each pixel p is obtained

$$\alpha_{ini}(p) = \arg\min_\alpha C_\alpha^{uocc}(p). \quad (10)$$

IV. Depth Regularization

In this section, we will show how to find the occlusion map and regularize the depth with a global energy function.

A. Occlusion Boundary Detection

We find the occlusion point using the visible constraint, i.e., the occlusion point is visible in the reference view and invisible in other views. In other words, if the difference of disparities of two neighboring pixels is larger than one pixel, there is an occlusion point here. In a light field, this constraint can only find the occlusion point in the central view. The threshold value ϵ_{occ} of the difference of disparities ought to be relaxed to find the occlusion point in other views,

$$\epsilon_{occ} = \frac{1}{\lfloor N_{uv}/2 \rfloor}, \quad (11)$$

where N_{uv} is angular sampling number of light field.

Since the initial depth of occluded point is unreliable and sometimes it is too smooth and random to distinguish the occlusion point accurately from its two neighbors, we select a disparity patch centered at each candidate occlusion point and use the K-means clustering to divide the patch into two classes (the feature is initial disparity). Then the difference of disparities of each class $\Delta\alpha$ is determined by,

$$\Delta\alpha = \left| \frac{1}{\alpha_1} - \frac{1}{\alpha_2} \right|, \quad (12)$$

where α_i is the center of the i-th class.

Finally the candidate occlusion point p is determined by comparing $\Delta\alpha(p)$ with the threshold ϵ_{occ},

$$Occ(p) = \begin{cases} 1 & \Delta\alpha(p) \geqslant \epsilon_{occ} \\ 0 & \Delta\alpha(p) < \epsilon_{occ}. \end{cases} \quad (13)$$

B. Un-Occluded Views Reselection

There are two drawbacks in the initial un-occluded views selection. Firstly, the patch size is an arbitrary value, which may lead to wrong estimation for occlusion types. Too small patch size leads multi-occluder occlusion becoming a single-occluder one. Conversely, too large size makes single-occluder occlusion being a multi-occluder one. A wrong occlusion type results in a wrong depth estimation. Thus, the patch size is an important factor to determine occlusion types. Secondly, some occlusion points are not detected using Canny detector in Section III-A, but their neighboring occluder pixels are detected (see Fig. 7). It is possible to recall these undetected occlusion points using an inverse un-occluded views selection after a voting stage. A complete opposite un-occluded views selection leads to wrong results which are wrong and cannot be estimated.

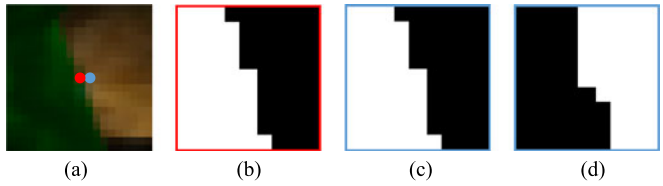


Fig. 7. The inverse un-occluded views selection. (a) is the scale-up of the Mona dataset [28], where the left green areas are the leaves and others are the background. (b) and (c) are the selected un-occluded views (white areas) for red and blue points in (a) after Section III-A. (d) is the ground truth of un-occluded views selection for the blue point in (a). It is noticed that the selected un-occluded views in (c) are nearly the inverse of the ground truth (d). Although the red point is an edge point, it is not an occlusion point and belonging to the leaves. But the K-means clustering selects the left areas as the un-occluded views. Since the blue point is not detected as edge point using Canny detector, its un-occluded views are voted from the red point. However, the blue point belongs to the background and the un-occluded views selection is inverse.

Based on the above analysis, the un-occluded views ought to be reselected. For each occlusion point p, we apply the K-means clustering for its local depth patch to find the background depth α_0 and occluder depth α_1, and the projection radius r is determined using (7). Then, each patch is resized to angular sampling number of light field. The following procedure for obtaining Ω_p^{uocc} is the same as Section III-A.

Since the occlusion points cannot be fully detected in Section IV-A, we retain the un-occluded views selection for other candidate points obtained in Section III-A.

C. Final Depth Regularization

In this step, given occlusion cues, we regularize the depth with Markov Random Field (MRF).

$$E = \sum_p E_{uocc}(p, \alpha(p)) + \lambda_s \sum_{p,q} E_{smooth}(p, q, \alpha(p), \alpha(q)), \quad (14)$$

where $\alpha(p)$ is the depth of pixel p. p and q are neighboring pixels. λ_s (set as 0.35 in the following experiment) is a factor to control the smoothness term.

The data term E_{uocc} measures the photo-consistency in the un-occluded views,

$$E_{uocc}(p, \alpha(p)) = 1 - e^{-\frac{C_{\alpha(p)}^{uocc}(p)^2}{2\sigma_{uocc}^2}}, \quad (15)$$

where σ_{uocc}(set as 3 in experiments) controls the sensitivity of the function to large distances. The definition of $C_{\alpha(p)}^{uocc}(p)$ can be found in (9).

Algorithm 1: Anti-Occlusion Depth Estimation Algorithm in Light Field.

Input:
4D light field LF
Output:
Final depth map α^*
Process:

$I_e = Canny(I)$	▷ Section III-A
$\Omega^{uocc} = Kmeans(I_e, I)$	▷ Section III-A
$\alpha_{ini} = InitialDepthEst(LF, I_e, \Omega^{uocc})$	▷ Section III-B
$Occ = OccDetection(\alpha_{ini}, I_e)$	▷ Section IV-A
$\Omega^{uocc} = Reselection(\alpha_{ini}, Occ)$	▷ Section IV-B
$E = E_{uocc}(LF, \Omega^{uocc}) + E_{smooth}(I, I_e, Occ)$	▷ Section IV-C
$\alpha^* = \arg\min_\alpha E$	▷ Section IV-C

The smooth term E_{smooth} encodes the smoothness constraint between two neighboring pixels,

$$E_{smooth}(p, q, \alpha(p), \alpha(q)) = \omega_{pq}|\alpha(p) - \alpha(q)|,$$
$$\omega_{pq} = e^{-\frac{(Occ(p)-Occ(q))^2}{2\gamma_{occ}^2} - \frac{(I_e(p)-I_e(q))^2}{2\gamma_e^2} - \frac{(I(p)-I(q))^2}{2\gamma_c^2}}, \quad (16)$$

where I_e is the edge map of central view image I. γ_{occ}, γ_e and γ_c are three weighting factors (set as 1.6, 0.8 and 0.08 respectively in experiments). Comparing with previous works, we introduce the occlusion term and edge term into the weighting function ω_{pq} to preserve sharp occlusion boundaries and keep the depth of occlusion boundaries similar.

The final depth map is obtained by minimizing

$$\alpha^* = \arg\min_\alpha E \quad (17)$$

using the GCO package [29]–[31].

The full description of the proposed algorithm is given in Algorithm 1. First, edge detection is applied on the central view image I to find all possible occlusion points I_e. Then, the un-occluded views Ω^{uocc} of each candidate occlusion point are selected based on a K-means clustering. After that, an initial depth map α_{ini} is estimated by using the un-occluded views. Moreover, the occlusion map Occ is detected by using the initial estimation. For each occlusion point, the un-occluded views are reselected according to the initial depth map. Finally, based on the un-occluded views and occlusion map, the depth map α^* is regularized with a MRF energy function.

V. Experimental Results

The results of our algorithm in different stages are shown in Fig. 8. Without the disparity prior, the projected radius in the first un-occluded views selection stage is not correct. The error leads to over-smoothed object boundaries in the initial depth map (Fig. 8(c)). Then, the occlusion map is detected with the help of initial depth. Finally the un-occluded views are reselected for depth regularization. It can be seen that the depth based reselection yields sharper object boundaries (Fig. 8(e)). Also the smoothness term decreases the noise to a large extent while maintaining the occlusion boundaries.

We compare our results with globally consistent depth labeling (GCDL) by [19], line assisted graph cut (LAGC) by [20], depth from combining correspondence and defocus cues (DFC) by [22], Bilateral Consistency Metric (BCM) by [21], phase-shift based depth estimation (PSD) by [23] and occlusion-aware depth estimation (OADE) by [25]. Noting that the results of GCDL, DFC, LAGC, PSD and OADE are obtained by running their public codes or executable files, and the results of BCM are provided by the authors. Due to the lack of source code, the BCM is only compared on synthetic datasets.

The performance of the proposed algorithm is evaluated on both synthetic and real light field datasets. We use light field benchmark dataset (LFBD, [28]) as synthetic data, which contains 7 light fields and each data includes a 9×9 light field and its ground truth depth. The real light fields contain two public datasets: (1) light field captured from Lytro Illum [32] which has a narrow baseline, (2) light field captured from a moving DSLR camera [33] which has a wider baseline.

A. Synthetic Datasets

Since the ground truth depth is provided in synthetic dataset, we provide complete evaluations on the un-occluded views selection, occlusion boundaries and depth map accuracy.

1) The Un-Occluded Views Selection: A consensus on depth estimation in computer vision is that more effective views lead to more accurate depth estimation. As a result, the precision and recall of the selected un-occluded views are important. We compute the F-measure[3] of the un-occluded views in occlusion using our algorithm and compare it with previous work [25].[4] The quantitative comparisons are listed in Table II and qualitative comparisons are shown in Fig. 9. The term 'Initial Selection' refers to initial result of un-occluded views selection. The 'Reselection' refers to the un-occluded views selection after Section IV-B.

In Table II and Fig. 9, our algorithm can always select more accurate un-occluded views than [25], and this advantage is more obvious in multi-occluder areas. However, [25] always selects more occluded views, which leads to over smooth results in occlusion areas (as shown in Figs. 11–13). The only exception is from Horses dataset, because there are many textures near occlusion boundaries in the background, and the color based K-means clustering cannot distinguish the background and occluder accurately in the complex texture areas. This issue will be further analyzed in Section V-D.

The reselected un-occluded views outperform the initial selection since the occlusion points are correctly detected in Section IV-A, and the projection radius are re-calculated according to the initial depth estimation.

[3]The harmonic mean of precision and recall compared with the ground truth

[4]It is noticed that although [21] first proposed to select the un-occluded views, the selected un-occluded views change with the change of refocus levels. So they cannot be predicted before the refocusing operation. [21] is hard to be compared and is not listed in the Table II. [25] first explicitly modeled the un-occluded views selection.

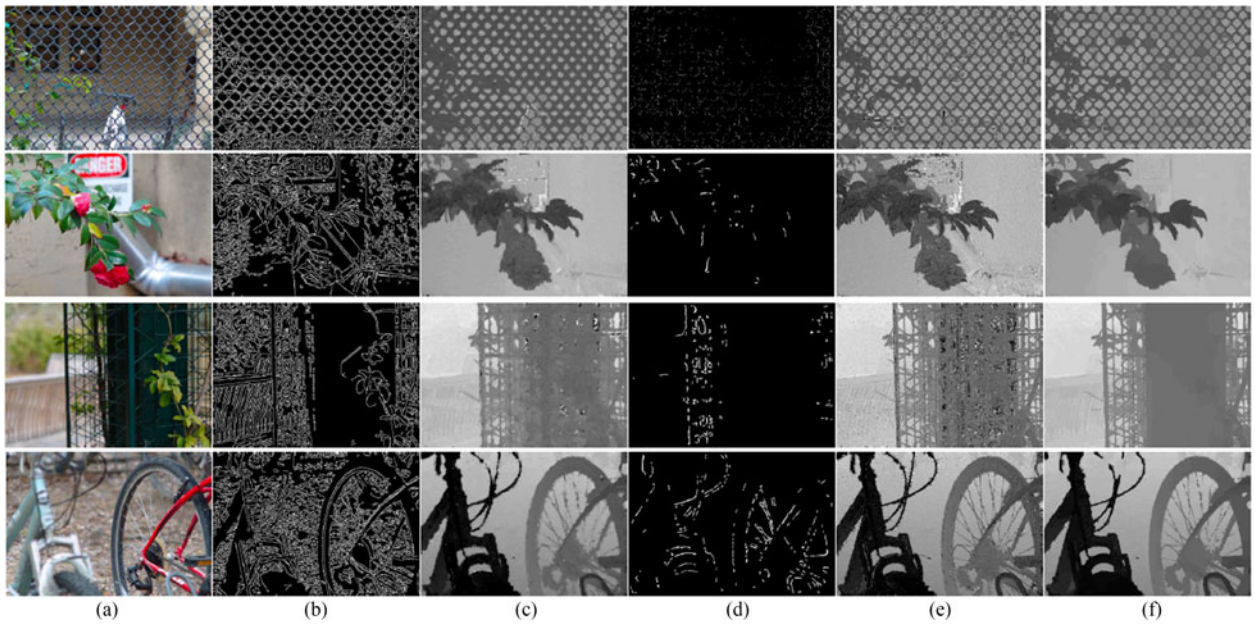


Fig. 8. The results of our algorithm in different stages on real light field data. (a) Input light field. (b) Edge detection. (c) Initial depth. (d) Occlusion detection. (e) Final depth w\o regularization. (f) Final depth with regularization.

TABLE II
THE F-MEASURE OF UN-OCCLUDED VIEWS SELECTION

	Buddha	Buddha2	Horses	Medieval	Mona	Papillon	StillLife	Average
OADE [25]	**0.71**	0.73	**0.66**	0.58	0.69	0.59	0.59	0.65
Initial Selection	0.67	0.66	0.60	0.57	0.75	**0.79**	0.68	0.67
Reselection	**0.71**	**0.74**	0.62	**0.60**	**0.79**	**0.79**	**0.72**	**0.71**

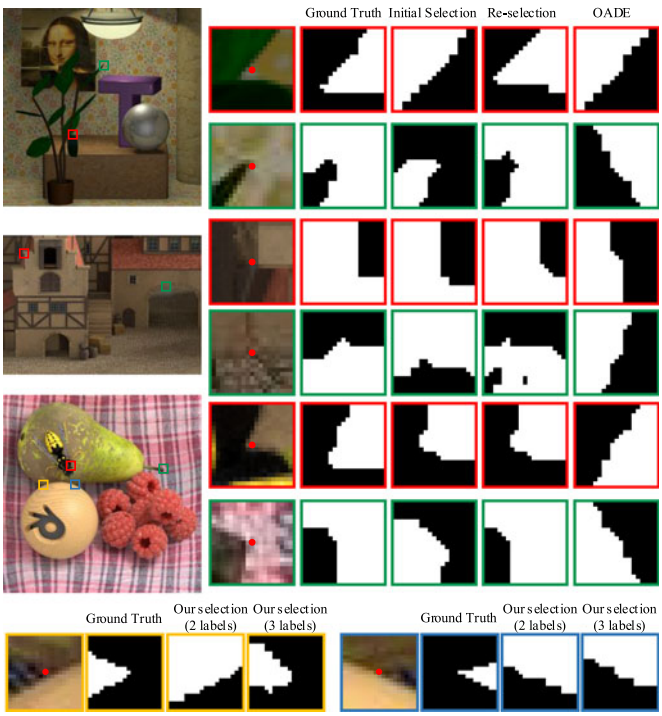


Fig. 9. Comparisons of un-occluded views selection. White areas denote the selected un-occluded views.

Since only two depth layers are used to model the occlusion in Section II, we also verify the performance of the method for multiple occluders at different depths, as well as depth-varying occluders. In the bottom row of Fig. 9, there are three depth layers, i.e., the ball, the pear and the tablecloth. Since there are two labels in the K-means clustering, the proposed method cannot accurately select the un-occluded views. This problem can be partly handled by increasing the number of labels (the orange boxes in Fig. 9), but it will increase the running time and is not always effective (it fails when the occluder shares the similar color with the background; see the blue boxes in Fig. 9). In some cases, it is sufficient to use two labels. Fig. 11(b) demonstrates that the reconstructed depth is fine for such cases. Although some of occluded views are improperly selected, our algorithm still achieves best results.

2) Occlusion Boundaries: For each light field, we detect occlusion boundaries using the depth map and compute its F-measure. Then, we compare it with other state-of-the-art algorithms. The quantitative comparisons are listed in Table III and the qualitative comparisons are shown in Fig. 10. Experimental results show that our algorithm outperforms previous works. Noting that the results of BCM [21] are not contained as it is difficult to run their codes in our experimental environment. However, as previous work [25] has demonstrated their advantages over [21], these comparisons are convincing. Our

TABLE III
THE F-MEASURE OF OCCLUSION BOUNDARIES

	Buddha	Buddha2	Horses	Medieval	Mona	Papillon	StillLife	Average
GCDL [19]	0.64	0.54	0.39	0.25	0.62	0.50	0.38	0.47
LAGC [20]	0.54	0.41	0.55	0.32	0.64	0.53	0.49	0.50
PSD [23]	0.62	0.79	0.60	0.48	0.76	0.53	0.68	0.64
OADE [25]	0.71	0.70	0.75	0.47	0.75	0.65	**0.82**	0.69
Ours	**0.75**	**0.85**	**0.80**	**0.56**	**0.81**	**0.76**	0.71	**0.75**

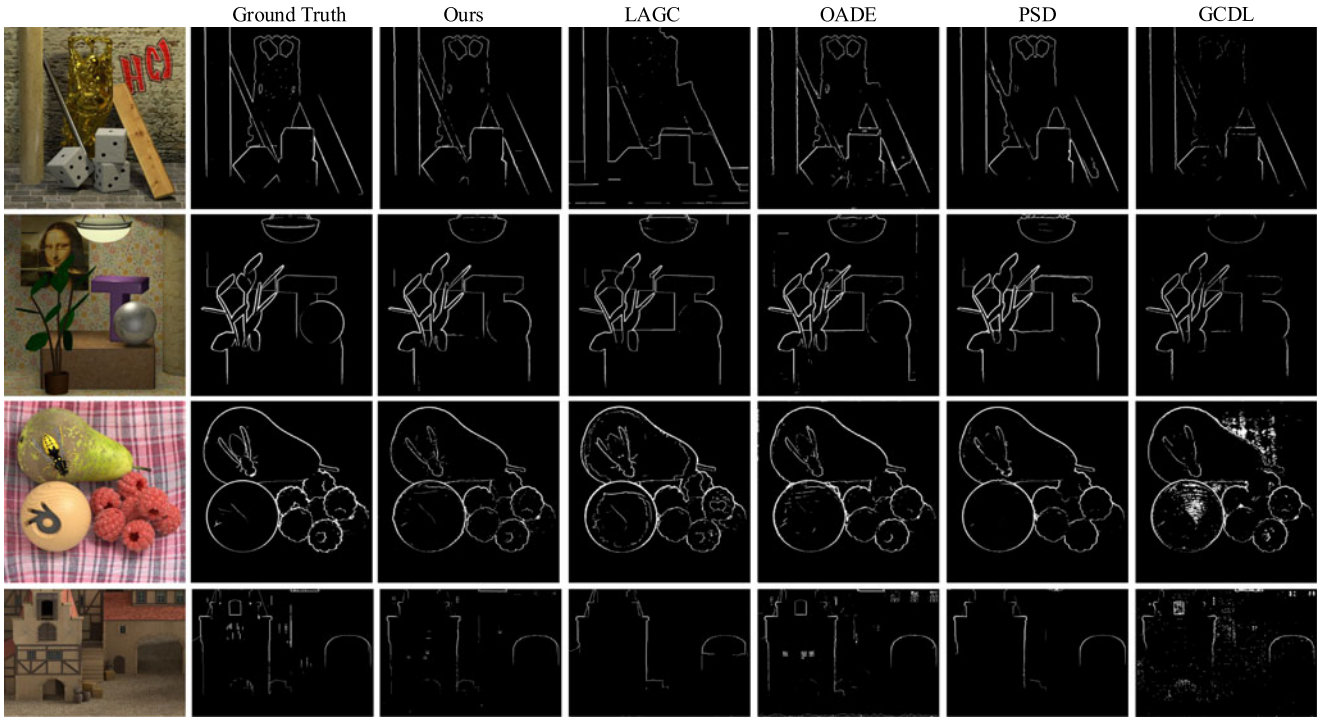


Fig. 10. The detected occlusion boundaries.

method performs not well in StillLife dataset (the third row in Fig. 10) because there are many weak occlusions (the difference of disparities is small) in StillLife dataset. The difference of disparities in the boundaries of the bee is small and the occlusion detection method in Section IV-A cannot handle these occlusions well.

3) Depth Maps: The quantitative comparisons of the RMS errors of the recovered disparity maps are listed in Table IV. Note that all results of our algorithm are obtained in a same parameter setting. Our algorithm outperforms state-of-the-art algorithms on almost all datasets. Apart from this, the RMS errors of initial results are also listed in Table IV, showing advantages of the un-occluded views reselection and regularization.

The qualitative comparisons of the estimated disparity maps are shown in Fig. 11. It can be seen that the GCDL and DFC always provide over-smooth results in occlusion boundaries and the reconstructed structures are thicker than the ground truth. The LAGC performs well for the thin structures (e.g., the branches and chandelier in Fig. 11(a)), however it fails when the structure boundary becomes short and incoherent (such as the red fruit in Fig. 11(b) and the ponytail in Fig. 11(d)). The PSD performs well for some occlusion areas (the branches in Fig. 11(a)), but it cannot provide a general solution for occlusions due to the lack of explicit occlusion model. The OADE performs well for single-occluder areas, however, the reconstructed occlusion boundaries are over-smoothed with the increase of occlusion boundaries (Fig. 11). Compared with these state-of-the-art algorithms, our approach yields sharper occlusion boundaries since our selection method can always find the un-occluded views accurately (shown in Fig. 9 and Table II) and does not select any other occluded views. Our results are sharper in the multi-occluder areas.

It is noticed that the proposed algorithm performs not as well as OADE in the un-occluded views selection on Horses dataset, however we can achieve better results on depth map. This is because our smoothness term can preserve occlusion boundaries better and the unreliable depths in some occlusion boundaries are smoothed by its neighboring occlusion points. Moreover, we obtain the best result on depth estimation for StillLife dataset, however the F-measure of detected boundaries is not the best. The reason is that our algorithm performs much better than OADE in multi-occluder boundaries where the difference of

TABLE IV
RMS ERRORS OF THE ESTIMATED DISPARITY FOR ALL PIXELS

	Buddha	Buddha2	Horses	Medieval	Mona	Papillon	StillLife	Average
GCDL [19]	0.079	0.094	0.163	0.111	0.096	0.158	0.184	0.126
DFC [22]	0.103	0.102	0.119	0.133	0.127	0.273	0.198	0.151
LAGC [20]	0.134	0.179	0.188	0.144	0.119	0.406	0.150	0.166
BCM [21]	**0.057**	0.139	0.122	0.129	0.077	**0.108**	0.113	0.106
PSD [23]	0.094	0.070	0.129	0.123	0.081	0.237	0.135	0.124
OADE [25]	0.095	0.107	0.140	0.115	0.089	0.125	0.212	0.126
Ours initial	0.124	0.117	0.198	0.111	0.136	0.229	0.226	0.163
Ours final	0.069	**0.051**	**0.074**	**0.101**	**0.071**	0.148	**0.110**	**0.089**

disparities is large and plays more important roles in the statistics of RMS errors.

However, our algorithm cannot handle the situation where the background has a similar texture or color to the occlusion. In Fig. 11(b) (green box), as the color of cloth in the background is similar to the foreground, it is difficult to recover the true depth. Apart from this, our algorithm cannot handle textureless areas just like all previous methods.

B. Real Datasets

To better verify the performance of the proposed algorithm on real light field data, we conduct two experiments. In the first experiment, light fields with a narrow baseline captured from a Lytro Illum camera are used. In the second experiment, light fields with a wider baseline captured from a moving DSLR camera are utilized. The disparities of two datasets range from -1.1 to 1.1 and 0 to 15 respectively. Since there is no ground truth in real light field data, we only perform qualitative analyses and show visual results.

1) Narrow Baselines: Figs. 12 and 13 show the comparisons of different algorithms. We replace the ground truth depth with the depth map provided by Lytro Illum software. It can be seen that there are many heavy occlusion areas in these datasets, including single-occluder and multi-occluder occlusions. The LAGC, PSD, GCDL and DFC fail in these datasets since the occlusion is not explicitly modeled or their cues are not robust enough to handle these heavy occlusion situations. In a rough looking, the results from Illum software and OADE achieve similar results with ours. To better illustrate the advantages of our algorithm, we scale up some heavy occlusion areas and the corresponding reconstructed depth maps of Illum software, OADE and ours. For these occlusion areas, our algorithm can always preserve the occlusion boundaries well, however, these areas are sometimes missing or over-smoothed in the results provided by Illum software and OADE.

For example, there are two flowers in the blue box of Fig. 12(a). Only our algorithm clearly preserves their boundaries and reconstructs the gaps between them, but these two independent flowers are connected into one in the results of Illum software and OADE (There are two independent branches in the green box of Fig. 12(c) and these structures are clearly reconstructed in our results. However, they are either missed or over-smoothed by the background (or themselves) in the Illum software and OADE (also see Fig. 13(a) and (b)). In Fig. 13(c), our algorithm reconstructs almost all spokes of the bike, however only a few of spokes are reconstructed in the results of Illum software and OADE.

2) Wider Baselines: Since the angular sampling of light field [33] is not square (the DSLR camera only moves along the horizontal axis.), the public codes of [19], [20], [22], [23], [25] are not applicable here. We only compare our results with [33]. Fig. 14 demonstrates comparisons of depth estimation results of two algorithms. Our algorithm achieves equal or better results than [33] in the Bike and the Couch. For example, the proposed algorithm reconstructs more spokes and basket than [33] in the Bike. However, we fail in the Statue and Mansion and some complex object boundaries are over-smoothed. The main reason is that light fields [33] are sampled along the horizontal axis. The proposed occlusion model is more suitable for square sampled light fields. Our algorithm cannot handle the top-down or approximate top-down occlusions in [33] since the occlusion boundaries are parallel to the sampling direction. In general, the proposed algorithm is more suitable for the narrow baseline and square sampling light fields.

C. Running Times

The comparisons of running time by different algorithms are listed in Table V. All algorithms are executed in a same desktop, with a 3.4 GHz CPU, 24G RAM and GTX 1080. The size of the input light field is $7 \times 7 \times 375 \times 541 \times 3$. The labels are set as 50 to balance the performance and the speed. To be fair, the programming language environments of different algorithms are also listed in Table V. The running time of the proposed algorithm is longer than most of baseline methods. Our current Matlab implementation consists of parallelized but un-optimized codes and we will optimize it in the future.

D. Limitation and Discussion

As mentioned above, the existing K-means strategy cannot handle specific case well, in which multiple occluders distribute in different depths (as illustrated in Fig. 9). Although more clustering labels can be set to improve the un-occluded views selection, it is time-consuming and is not always effective. It is essential to develop a more robust strategy for un-occluded views selection. Additionally, OADE performs better than ours for un-occluded views selection on Horses dataset (as

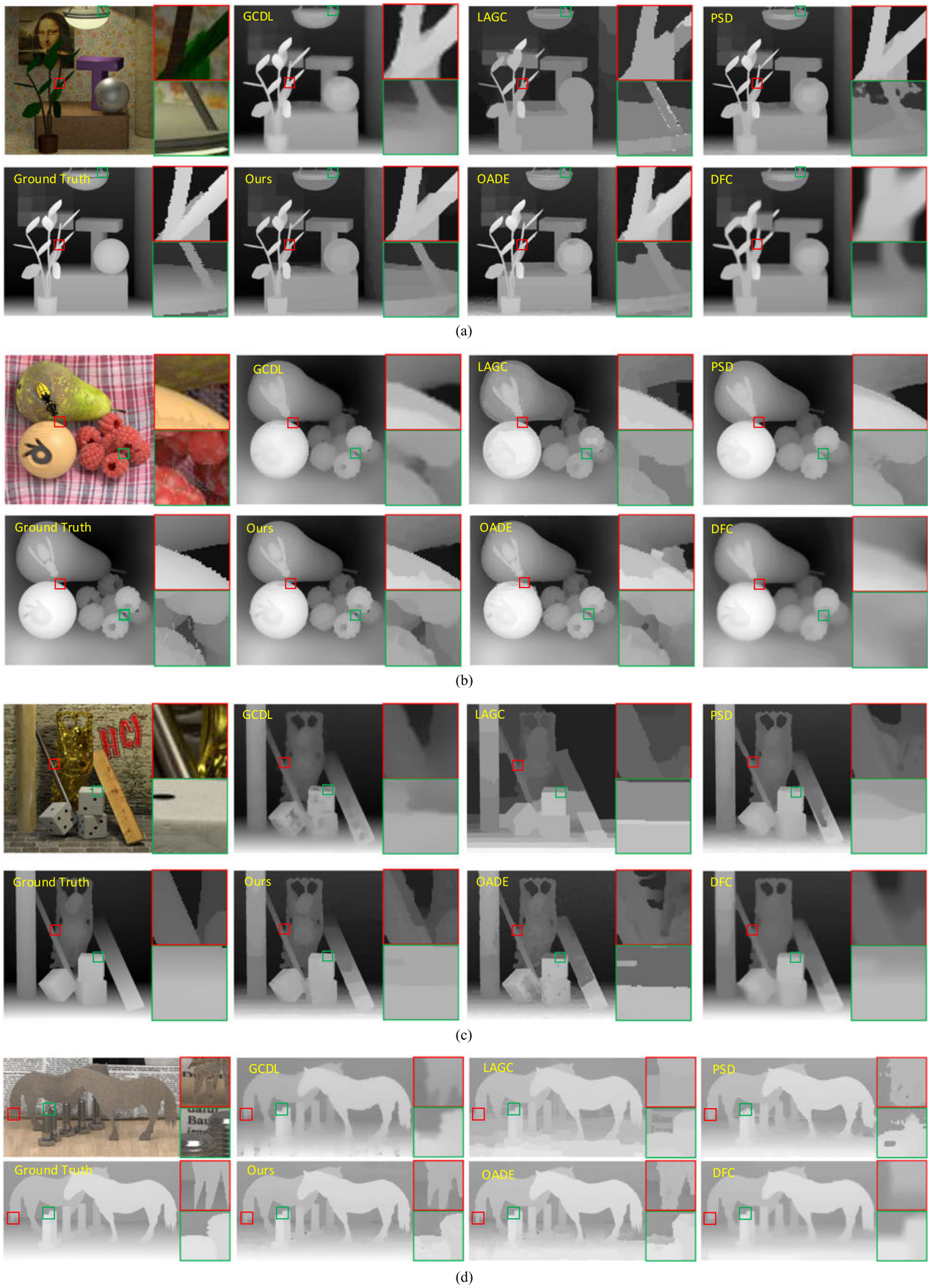


Fig. 11. The disparity maps of the Mona, StillLife, Buddha2 and Horses respectively. (a) Mona. (b) StillLife. (c) Buddha2. (d) Horses.

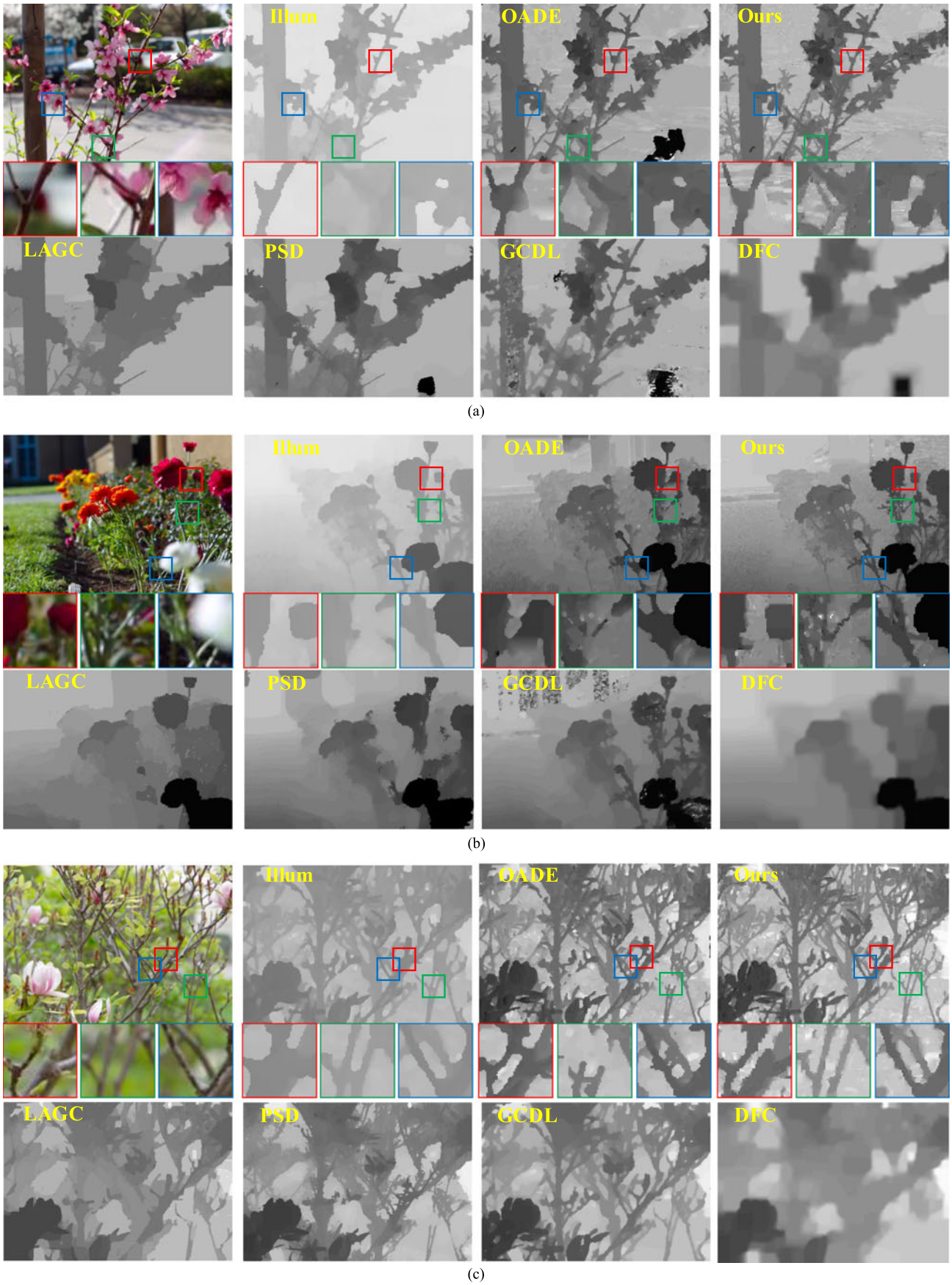
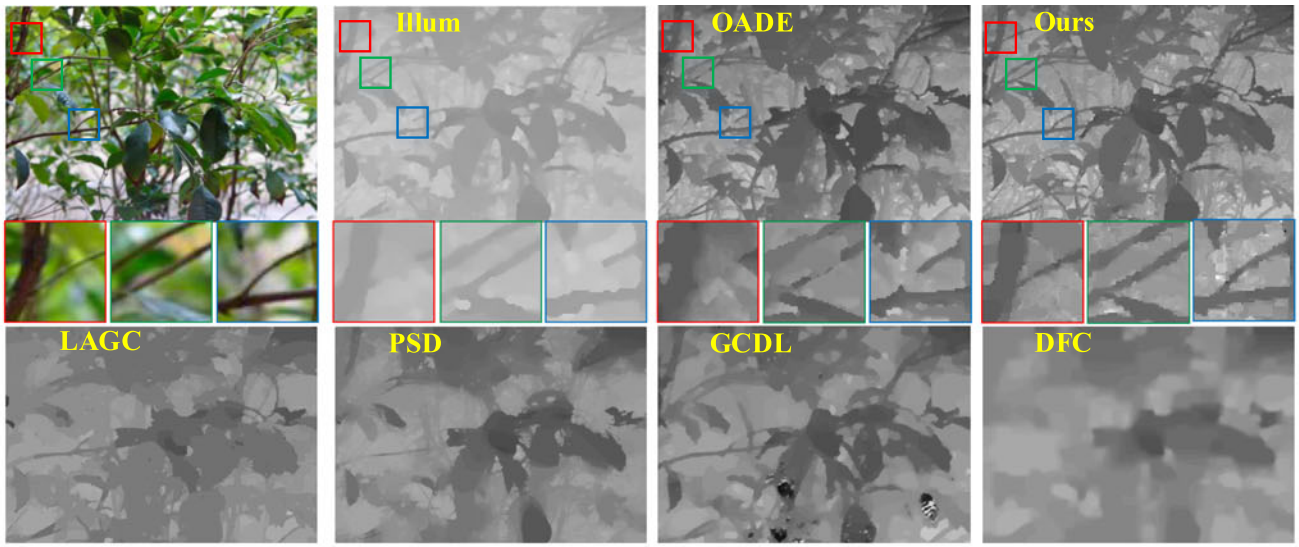
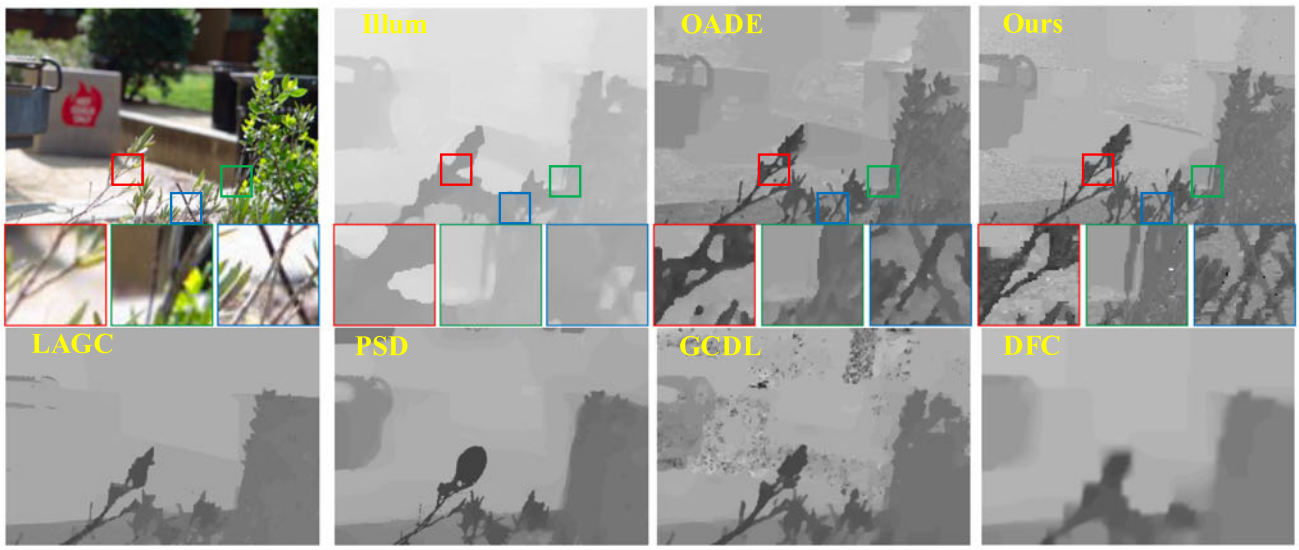


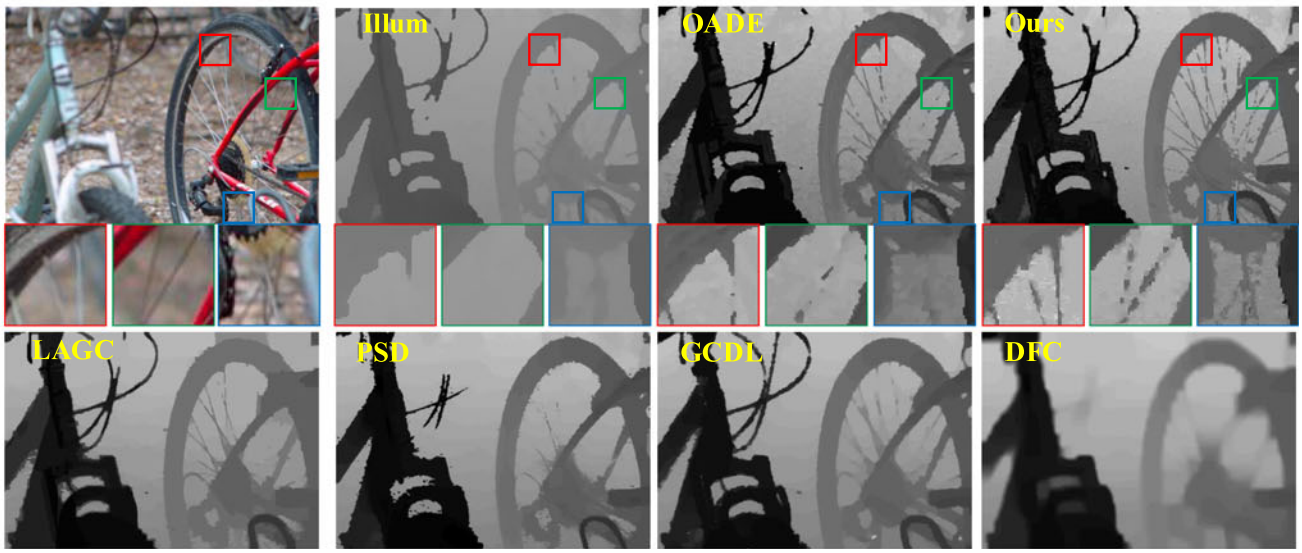
Fig. 12. The disparity maps of the Flowers1, Flowers2 and Flowers3 respectively. (a) Flowers1. (b) Flowers2. (c) Flowers3.



(a)



(b)



(c)

Fig. 13. The disparity maps of the Plants1, Plants2 and Bike respectively. (a) Plants1. (b) Plants2. (c) Bike.



Fig. 14. The disparity maps of the Bike, Couch, Statue and Mansion respectively. (a) Bike. (b) Couch. (c) Statue. (d) Mansion.

TABLE V
COMPARISONS OF RUNNING TIME BY DIFFERENT ALGORITHMS

	GCDL [19]	DFC [22]	LAGC [20]	PSD [23]	OADE [25]	Ours
Time	**48 s**	192 s	85 min	575 s	149 s	577 s
Language	C++/CUDA	Matlab/C	C/C++	Matlab	Matlab/C	Matlab

shown in Table II). In Fig. 15, the boundary of occluded/un-occluded views ought to be along the boundary of horses. However, due to the complex textures in the background, there are three colors in the neighboring patch of the occluded point (the brown horses, the white paper and the black characters), our K-means based selection cannot properly select the un-occluded views.

Apart from this, our algorithm performs well for some shading areas. There is an illumination change between the above and front in the dice of Fig. 11(c), and highlights on the road belt in Fig. 12(a) (lower right). Compared with others, our algorithm accurately recovers the shape of these areas. This is a topic worthy discussing since the shading is not considered in our algorithm.

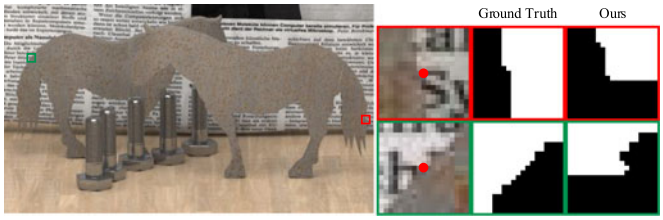


Fig. 15. The limitation of K-means based un-occluded views selection.

VI. Conclusion and Future Works

The multi-view nature is only one expression of 4D light field and there are many other representations, such as the EPI or focal stack [34]. Apart from the occlusion, it is also challenging to model other classical problems, such as shading, low-texture and repeatable pattern, in depth estimation using different representations of 4D light field. Wang *et al.* [35] did a good try, however there are more depth cues worthy to utilize in full 4D light field. We will focus on solving other problems in depth estimation using light field.

Acknowledgment

The authors would like to thank Q. Zhang for helpful discussions and anonymous reviewers and editors for the comments on the manuscript.

References

[1] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.
[2] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2806–2813.
[3] X. Zhang, Y. Wang, J. Zhang, L. Hu, and M. Wang, "Light field saliency vs. 2D saliency: A comparative study," *Neurocomputing*, vol. 166, pp. 389–396, 2015.
[4] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3061–3070.
[5] V. Vaish, M. Levoy, R. Szeliski, C. L. Zitnick, and S. B. Kang, "Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 2331–2338.
[6] Z. Xiao, L. Si, and G. Zhou, "Seeing beyond foreground occlusion: A joint framework for SAP based scene depth and appearance reconstruction," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 979–991, Oct. 2017.
[7] A. Jarabo, B. Masia, A. Bousseau, F. Pellacini, and D. Gutierrez, "How do people edit light fields?" *ACM Trans. Graph.*, vol. 33, no. 4, pp. 146-1–146-10, 2014.
[8] H. Mihara, T. Funatomi, K. Tanaka, H. Kubo, Y. Mukaigawa, and H. Nagahara, "4D light field segmentation with spatial and angular consistencies," in *Proc. IEEE Int. Conf. Comput. Photography*, 2016, pp. 1–8.
[9] Z. Zhang, Y. Liu, and Q. Dai, "Light field from micro-baseline image pair," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3800–3809.
[10] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, 2016, Art. no. 193.
[11] Z. Xiao, Q. Wang, G. Zhou, and J. Yu, "Aliasing detection and reduction scheme on angularly undersampled light fields," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2103–2115, May 2017.
[12] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Proc. 7th Eur. Conf. Comput. Vis.*, 2002, pp. 82–96.
[13] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon, "Global stereo reconstruction under second-order smoothness priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2115–2128, Dec. 2009.
[14] M. Bleyer, C. Rother, and P. Kohli, "Surface stereo with soft segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1570–1577.
[15] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*. Cambridge, MA, USA: MIT Press, 1991, pp. 3–20.
[16] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Ann. Conf. Comput. Graph. Interact. Tech.*, 1996, pp. 31–42.
[17] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. 23rd Ann. Conf. Comput. Graph. Interact. Tech.*, 1996, pp. 43–54.
[18] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Stanford Univ., Stanford, CA, USA, CSTR Tech. Rep. 2005-02, 2005.
[19] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4D light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 41–48.
[20] Z. Yu, X. Guo, H. Lin, A. Lumsdaine, and J. Yu, "Line assisted light field triangulation and stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2792–2799.
[21] C. Chen, H. Lin, Z. Yu, S. Kang, and J. Yu, "Light field stereo matching using bilateral statistics of surface cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1518–1525.
[22] M. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 673–680.
[23] H.-G. Jeon *et al.*, "Accurate depth map estimation from a lenslet light field camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1547–1555.
[24] H. Tang and K. N. Kutulakos, "What does an aberrated photo tell us about the lens and the scene?" in *Proc. IEEE Int. Conf. Comput. Photography*, 2013, pp. 1–10.
[25] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3487–3495.
[26] R. Ng, "Digital light field photography," Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, 2006.
[27] Y. Anzai, *Pattern Recognition & Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2012.
[28] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," in *Proc. Annu. Workshop Vis., Model. Vis.*, 2013, pp. 225–226.
[29] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.
[30] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[31] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.
[32] A. S. Raj, M. Lowney, and R. Shah, "Light-field database creation and depth estimation," 2016.
[33] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, "Scene reconstruction from high spatio-angular resolution light fields." *ACM Trans. Graph.*, vol. 32, no. 4, pp. 73-1–73-12, 2013.
[34] A. Levin and F. Durand, "Linear view synthesis using a dimensionality gap light field prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1831–1838.
[35] T.-C. Wang, M. Chandraker, A. Efros, and R. Ramamoorthi, "SVBRDF-invariant shape and reflectance estimation from light-field cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5451–5459.

Author's photographs and biographies not available at the time of publication.

Seeing Beyond Foreground Occlusion: A Joint Framework for SAP-Based Scene Depth and Appearance Reconstruction

Zhaolin Xiao, Lipeng Si, and Guoqing Zhou

***Abstract*—Foreground occlusion is a significant challenge in three-dimensional (3-D) reconstruction. In this paper, we first characterize the differences between multiview reconstruction with and without foreground occlusion. In order to reconstruct target scene, we attempt to see through the foreground occlusion using synthetic aperture photography. Different from existing methods, we propose a more generalized model for scene reconstruction, in which the target scene may not be fully observed by any of reference view. Assuming both scene depth and appearance are unknown, we reconstruct 3-D scene from camera array data by selecting optimal views with pixel based clustering. Then, we propose an iterative reconstruction approach in global optimization framework, in which we refine the reconstruction results by applying a coarse-to-fine strategy. Even when all views are partially occluded, our approach can recover accurate depth map as well as scene appearance using camera array data. Experimental results have indicated that the proposed approach is more robust to foreground occlusions and outperforms state-of-the-art approaches.**

***Index Terms*—Camera array, depth recovery, light field, occlusion handling, scene reconstruction, synthetic aperture photography.**

I. Introduction

IN EVERYDAY environments, human beings can rapidly recognize a specific object with our vision system, even when the object is partially occluded. Evidence from neurosciences in humans and human-like animals suggests that this kind of occlusion handling ability involves extra-striate visual cortex, which is described as a retinotopic homologue of ventral V4 area in human brain [1]. Moreover, a posterior parietal cortex may be involved in processing occlusions. From fMRI observation data, researchers proved posterior parietal cortex is not only sensitive to spatial occlusions, but also sensitive to temporal occlusions [2]. This means that, when inferring an occluded object, human brain is seeking more spatiotemporal information than a pair of binocular images. Unfortunately, expressing the occlusion handling ability in a mathematical way is still beyond the confines of human knowledge.

With rapid advances in computer hardware and visualization systems, computer vision is creeping into nearly every corner of science. Reconstructing three dimension geometry and photometry of real world scene from two or more images remains a difficult problem in computer vision. Due to the complexity of nature world, scene reconstruction can mainly be deteriorated by depth discontinuities, textureless regions, noises and occlusions [3], [4]. In this paper, we focus on the occlusion-handling problem in image based scene reconstruction, especially for the foreground occlusion cases, as shown in Fig. 1(a).

Anti-occlusion scene depth estimation is one of the greatest challenges in binocular stereo and multiview stereo (MVS). One straightforward method is applying cross-checking using uniqueness constraint [5], which can detect the occluded image region efficiently. Zitnick *et al.* [6] proposed an iterative algorithm to estimate the occlusion in a 3D array, which also used the uniqueness constraint. Geiger *et al.* [7] suggested a monotonicity constraint to keep the ordering of matching pixels. To handle occlusion for more general scenes, Sun *et al.* [8] proposed a symmetric stereo model by using visibility constraint. Since the valid matching is not guaranteed for each pixel in binocular stereo, traditional stereo schemes tend to introduce holes in the result [9]. One solution is to detect and model the occlusion explicitly, as suggested in [8], [10], [11]. Another solution is to perform global optimization to smooth the results reasonably, such as graph-cuts [12]–[14], belief propagation [15]–[17] and dynamic programming [18]–[20]. To further explore the performance of these methods, we recommend to refer [21], which compares different anti-occlusion algorithms in a systematic way.

It is worthwhile to note that, an involuntary assumption is implicitly shared in previous work. There is always a reference image or view, in which the target scene is completely visible. For simplicity, we call the existence assumption of a occlusion free reference image as the reference-existence-condition (REC) assumption in this paper. Since the objects may frequently overlap each other, the REC assumption will hardly be satisfied even in moderately complicated scenes. For example, seeing a yard through fences, our visual system can easily filter out the occlusion from the target scene. To solve this spatial coherence between occlusion and target, we would prefer to slightly switch our view point rather than to stare the target with a static view. To carefully review this process, the REC assumption is apparently

Manuscript received December 30, 2016; revised April 6, 2017 and May 20, 2017; accepted June 7, 2017. Date of publication June 12, 2017; date of current version October 23, 2017. The work was supported in part by NSFC funds (61501370, 61531014, and 61401359), in part by a Research Grant of ShaanXi Province (2016JQ6069), and in part by a Research Grant of the State Key Laboratory of Virtual Reality (BUAAVR-16KF-10), China. The guest editor coordinating the review of this paper and approving it for publication was Prof. Yebin Liu. *(Corresponding author: Zhaolin Xiao.)*

Z. Xiao is with the Department of Computer Science, Xi'an University of Technology, Xi'an 710048, China (e-mail: zlin.xiao@gmail.com).

L. Si and G. Zhou are with Northwestern Polytechnical University, Xi'an 710065, China (e-mail: silp@live.cn; zhouguoqing@nwpu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2715012

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.



Fig. 1. Reconstructing scene behind foreground occlusion. (a) A scene with complex foreground occlusion. (b) Reconstructed depth map. (c) Recovered appearance.

failured in this visual task, but our brain can also reconstruct the target scene with proper spatiotemporal information.

But so far, little research has been involved in explicitly modeling scene reconstruction when the REC assumption is violated. In this paper, we propose a novel iterative approach based on camera array data, which can estimate both depth and appearance of a scene behind significant foreground occlusions, as shown in Fig. 1(b) and (c). We shows an overview of the proposed approach in Fig. 2. The major contributions of the paper include,

1) We characterize the differences between scene reconstruction with and without foreground occlusion, which leads to a more general reconstruction model.
2) We propose an iterative reconstruction approach in global optimization framework, which recovers both scene depth and appearance.

II. Related Work

Supposing the foreground occlusion existing in all views, e.g., in Fig. 1(a), the occlusion can hardly be separated or modelled explicitly, so that the ideal REC is not satisfied in this case. To handle occlusion problem, researchers tend to obtain more spatiotemporal information, which is similar to what human brain does.

A straight forward solution is first to apply occlusion active contours segmentation [22], [23], then to apply pixel-wise or patch-wise image inpainting [24], [25], which can refill the target region with a visually plausible approximation of image content. But image inpainting technique always need the region mask of occlusion as a prior, that is unknown in most of scene reconstruction applications.

Another effective solution is modifying imaging optics and processing of vision system.In general, any kind of imaging process can be treated as a downsampling of the 7D plenoptic function [26]. Levoy and Hanrahan [27] suggested that both positional and angular information of light rays can be parameterized by a pair of parallelled planes, which is known as the 4D light field. They had experimentally sampled a time invariant 4D light field by fixing one camera on a robot gantry. Further, Wilburn *et al.* [28] built a camera array system (128 cameras) to sample the time variant 4D light field. The compact size light field cameras such as Lytro [29] (the plenoptic 1.0) and Raytrix [30] (the plenoptic 2.0) are now available in the commodity market. Compare to traditional projective model based imaging technique, the light field imaging has two major advantages. One advantage is the interactive image generation ability, such as digital refocusing [31], viewpoint switching [28], depth of field extending [32]. The other advantage is the scene reconstruction ability, especially for scene with occlusion.

In the past decade, researchers have achieved some remarkable progresses in light field data based scene reconstruction. With variational Bayesian theory, Bishop and Favaro [33] proposed a depth reconstruction and image super-resolution iteration framework on the focused light field camera (known as the plenoptic 2.0), e.g., the Raytrix [30] camera data. Through light field angular signal analysis, Wanner and Goldluecke [34] proposed an epipolar plane image (EPI) based depth estimation algorithm. Different from multiview correspondence, on the EPI image, the slope of the line is regarded as a scene depth cue. Kim *et al.* [35] presented a more smooth reconstruction from high spatio-angular resolution light fields. Tao *et al.* [36] combined correspondence, defocus and shading cues in the 4D light field to reconstruct accurate scene depth. Jeon *et al.* [37] introduced a phase shift based sub-pixel-wise scene reconstruction algorithm, which can accurately estimate depth map from light field data. Lin *et al.* [38] further improved the reconstruction quality by introducing light field focal stack symmetry. However, the above methods are vulnerable in presence of heavy foreground occlusion.

To handle occlusion problem, Chen *et al.* [39] suggested a novel bilateral metric on light field angular patch, on which the occlusion view can be inferred by comparing to central view. Assumming photo consistency remains in angular patch, Wang *et al.* [40] separated angular patch into non-occlusion and occlusion parts respectively. Then, they introduce an occlusion aware regularization algorithm to implement accurate depth estimation. Williem and Inkyu [41] introduced an angular entropy metric and adaptive defocus response, which can be used as a robust data cost for depth estimation of noisy scene. It's important to notice that object appearance estimation is another goal for reconstructing the scene behind foreground occlusion, e.g., in point patterns classification [42] and multiple view analysis [43]. But in these work, the REC assumption is implicitly needed and occluded objects still only exist in non-reference views.

In order to enhance the visibility of imaging system, Levoy *et al.* [44] employed a video projector, a camera, and an array of mirrors to implement the confocal imaging, which can see through foreground occlusion and scattering medium, such as

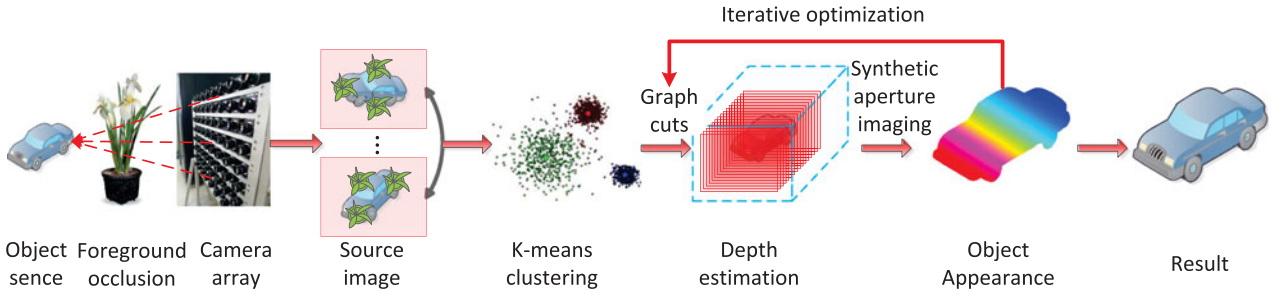


Fig. 2. An overview of the proposed approach.

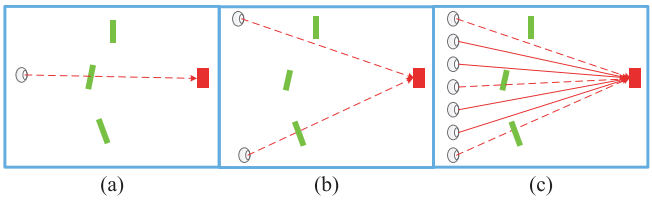


Fig. 3. Monocular, binocular and multi-view vision system with foreground occlusion. (a) Monocular vision system. (b) Binocular stereo vision system. (c) Multi-view stereo vision system.

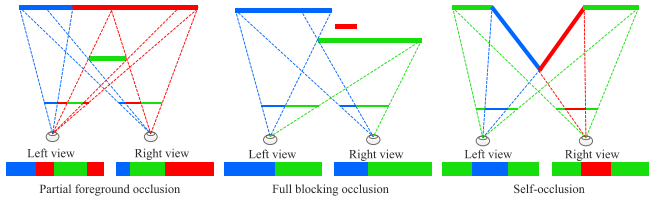


Fig. 4. An illustration of three different types of occlusions.

foliage, weakly scattering fog, murky water. Vaish *et al.* [45] presented an entropy based algorithm on camera array data, which can reconstruct the occluded surface from a set of partially occluded images. Do *et al.* [46] suggested an independent component analysis and K-means algorithm to apply 3D reconstruction of occluded objects. McCloskey [47] suggested a custom aperture mask for estimating the foreground occluder, so that one can recover the partially occluded scene with a modification of light field rendering. Pei *et al.* [48] suggested to suppress the foreground occlusion via pixel labeling technique, they first mark foreground image pixels by solving a binary labeling problem on source images, and then all the marked pixels are not contributed to the next traditional synthetic aperture imaging. Zhu *et al.* [49] derived a occluder-consistency constraint between light field spatial and angular spaces, which can be used as a guidance to select the un-occluded views, especially in multi-occluder cases. In this paper, we characterize a more general model for reconstructing both depth and appearance of target scene. Different from existing work, our model can well handle most of partially occluded cases, even when the REC assumption is untenable.

III. Problem Formulation

With a monocular camera, we can only sample and quantize the 2D luminous intensity distribution on the image sensor plane. Therefore, the angular information of incidence light rays is unrecoverably lost due to lens converging. By using a binocular stereo system, we can acquire two different angular samples of a space point, as shown in Fig. 3(b). Due to the lack of samples, most of monocular and binocular images based scene reconstruction algorithms would fail because of significant occlusions. MVS is naturally more robust than monocular or binocular stereo, as shown in Fig. 3(c).

Assuming that we estimate the depth map D of a natural scene from a collection of K view images $I = \{I_k\}$. A promising solution is inferring a "best guess" from I, which can be formulated as a *maximum a posteriori* (MAP) estimation. Therefore, the depth estimation becomes maximizing the probability of $P(D|I) = \frac{P(I|D)P(D)}{P(I)}$ by Bayesian law, $P(D|I) \propto P(I|D)P(D)$. In the Markov network based MVS, the probability maximization is equivalent to energy minimization, which can be solved by traditional global optimizations. In general, the energy function is modeled as,

$$E = E_{data} + E_{cons} \tag{1}$$

where E_{data} is the matching cost and E_{cons} is the constraint term, corresponding to $P(I|D)$ and $P(D)$ respectively.

Under the circumstance of above-mentioned REC, we have a hypothesis that the target scene is almost completely visible in the reference view I_{ref}. Suppose that the homography $H_k(d)$ represents transformation matrix which registers the image I_k into the coordinate system of reference view. The matching cost E_{data} is commonly represented as $\|H_k^{-1}(d)I_k - I_{ref}\|_2$, which reflects the matching consistency of different angular samples traced back at the given depth d. By temporally simplifying the constraint term E_{cons}, estimating depth map d^* can be roughly written as

$$d^* = \arg\min_d \sum_{k=1}^{K} \|H_k^{-1}(d)I_k - I_{ref}\|_2 \tag{2}$$

Since there is only one variable d in (1), a variety of specific scene priors have been proposed to improve the term E_{cons}, such as smooth constraint, ordering constraint, uniqueness constraint and visibility constraint [21]. But, as far as we know, very little discussion refers to what if the I_{ref} is unknown, or I_{ref} is deteriorated by some occlusions. Here, we can roughly categorize the occlusions into three different types, i.e., partial foreground occlusion, full blocking occlusion and self-occlusion, as shown in Fig. 4. The major concern of the paper is partial foreground occlusion handling. Distinct from self-occlusion and full blocking occlusion cases, target scene under partial foreground occlusion can be objectively separated by its depth. In partial

foreground occlusion, any parts of target scene can be visible in some specific views.

To put forward a new theoretical problem, we define the space of unknowns as $S(D, O, I_{ref})$, where D is depth map, O is undetectable partial foreground occlusion, and I_{ref} is the unknown reference image respectively. Then, our new MAP problem becomes to maximize the *posterior* probability of $P(S|I) = \frac{P(I|S)P(S)}{P(I)}$. The likelihood $P(I|S)$ depends on the imaging formation, in the presence of foreground occlusion. So we model this process as

$$I_k = B_k(O, d)H_k(d)I_{ref} \tag{3}$$

where $B_k(O, d)$ is a blocking function depending on the depth d, occlusion O and the view of camera k. Since usually the occlusion O is undetectable, it is impractical to estimate I_{ref} from any I_k directly. Here, to separate foreground occlusion, the only assumption is that we limit the target scene within a depth range $[d_a, d_b]$. For this reason, any scene elements has smaller depth than d_a is deemed as foreground occlusion, which corresponds to O in (3). This assumption is similar to forcing the disparity within a limited range, which is commonly used in some other papers.

Distinct from (2), we put forward a new model for scene estimation,

$$\langle d^*, I^*_{ref}\rangle = \arg\min_{d, \hat{I}_{ref}} \sum_{k=1}^{K-n} \|H_k^{-1}(d)\hat{I}_k - \hat{I}_{ref}\|_2 \times \rho(n)$$
$$\text{s.t. } d_a \le d \le d_b \tag{4}$$

where $\hat{I}_k$ is potentially occlusion-filtered-out view and $\hat{I}_k = B_k^{-1}(O, d)I_k$. $\hat{I}_{ref}$ denotes the occlusion-removed image of the reference view. Theoretically, our model is more general than traditional one in (2), in which only the scene depth is unknown. Furthermore, we allow to give up n matching candidates at the cost of $\rho(n)$ in the energy minimization. Once the foreground occlusion is taken out or may be detected, the blocking function will be treated as a constant, such that the model in (4) is degraded to traditional one in (2).

Please notice that, there are two major differences between reconstruction with and without foreground occlusion. First, with foreground occlusion, the ideal reference view I_{ref} in (2) is unknown, so both scene appearance $\hat{I}_{ref}$ and depth d need be simultaneously estimated in (4). Secondly, unlike (2) model, to reconstruct a specific point, the (4) model allow us to give up some views to suppress the negative effect of foreground occlusion. In the meantime, all existing constraints can still be hold in our model, such as uniqueness constraint, smooth constraint, visibility constraint, etc [5], [21], [50].

IV. Joint Energy Function for Scene Depth and Appearance Reconstruction

As we mentioned above that O is undetectable, we can not derive the blocking function $B_k(O, d)$ explicitly. For this reason, (4) cannot be directly used as a matching cost. Recalling that $\hat{I}_{ref}$ is the occlusion-removed image of the reference view, the different image blocks of ideal $\hat{I}_{ref}$ will be randomly distributed in those occluded views, and usually be accompanied by noises or outliers. On the contrary, in those non-occluded views, we have the fact that $\hat{I}_{ref} = I_{ref}$ since the blocking function is treated as a constant. Thus, those observed $\hat{I}_k$ s remain photo consistency at the correct depth. If we have enough images of non-occluded views, we can distinguish them from such strong occluded noise. In this case, it is possible to estimate depth without explicitly modeling occlusion O.

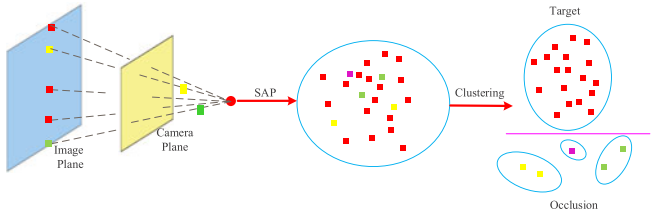


Fig. 5. Separating occlusion from the target by clustering.

Similar to volumetric reconstruction, we first divide target scene space into multiple depth layers. We initialize its constituent of pixels with synthetic aperture photography (SAP), by which one can adjust image focal plane after a single exposure on 4D light field [28]. If we have prior pixelwise depth map, we can directly integrate different light field angular samples to obtain scene appearance using (5). Unfortunately, since accurate scene depth D and occlusion dependent $I(d)$ are both unknowns in the initial stage, we can only apply a preliminary screening process on light field samples, as shown in Fig. 5.

$$I(d) = \frac{1}{K}\sum_{k=1}^{K} H_k^{-1}(d)I_k \tag{5}$$

To handle foreground occlusion, (4) allows us to pick up $K - n$ views selectively and minimize the contributions of the other unexpected occluded n views. Therefore, we propose a K-means based matching cost which considers both consistency and the number of involved views. Recalling the overview of the proposed approach in Fig. 2, after acquiring multiple views of the target scene, we compute the matching cost $F(p, d)$ for each pixel p at each depth layer d. Unlike existing approaches, we use a K-means algorithm [51] to pre-select matching candidates. In the paper, our input of K-means algorithm is a bunch of RGB image pixels from different angular views. Suppose that those I_k s are clustered into m groups, we select the group C_s that includes most of points, whose in-class distance is $dt_s = \frac{1}{ct_s}\sum_{I_k \in C_s} \|I_k - \mu_s\|_2$, where ct_s is the number of points in C_s and μ_s is the cluster centroid.

$$F(p, d) = \begin{cases} \dfrac{dt_s}{ct_s}, & \text{if } dt_s \le T \\ +\infty, & \text{Otherwise} \end{cases} \tag{6}$$

where T is a consistency threshold. The proposed $F(p, d)$ is encouraging to assume the cluster centroid μ_s as the appearance $\hat{I}_{ref}$ of pixel p at current depth layer d. While a 3D objective points do exists on layer d, there should be a cluster with small in-class distance dt_s and large number of in-class samples ct_s, so that the $F(p, d)$ can be very small. Whereas, a large $F(p, d)$ indicates lower possibility of the existence of any 3D objective point on layer d. Since we always choose the cluster with maximum number of samples, it is necessary to employ a consistency threshold T to guarantee samples originate from a same 3D objective point. In the paper, we calculate T as the average in-class

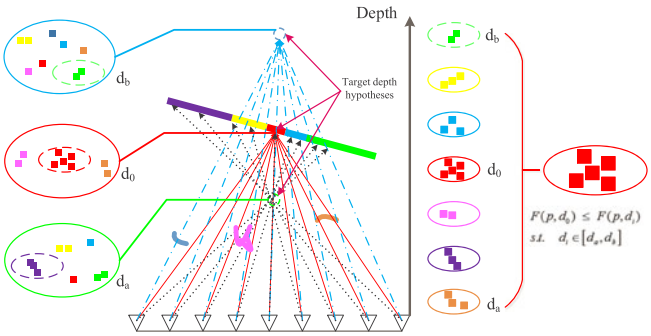


Fig. 6. Explanation of angular samples distribution on one point clustering case.

distance of all clustering groups, as below.

$$T = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{ct_i}\sum_{I_k \in C_i}\|I_k - \mu_i\|_2\right). \tag{7}$$

According to (4), we suppress the affects of foreground occlusion by giving up some views, which are more likely to be randomly distributed in clustering process. To better understand why the foreground occlusion affects can be suppressed by clustering, we intuitively explain a simple one point case of clustering in 2D sampling space in Fig. 6. A colored target scene around depth d_0 is captured by multiple cameras, which is partially blocked by some irregular-shaped occlusions. If the depth hypotheses of target are at different layers, e.g., at depth d_a, d_0 and d_b. Then, all corresponding angular samples at each depth are shown on the left of Fig. 6. At the correct depth d_0, the photo consistency of non-occluded view samples is still hold, while the occluded view samples are distributed randomly, because they may originate from different occlusion points. Whereas, considering the wrong depth hypotheses, e.g., at depth d_a, d_b, the corresponding angular samples may also appear random distribution, so that matching cost $F(p, d_i)$ will be larger than $F(p, d_0)$ at the correct depth, as shown on the right of Fig. 6.

In the paper, we employ the similar smooth constraint and neighborhood system $\mathcal{N}$ as described in [12]. The smooth cost $V_{p,q}(d(p), d(q))$ is defined as

$$V_{p,q}(d(p), d(q)) = \omega(p, q, \lambda)\min(L, |d(p) - d(q)|) \tag{8}$$

where L is a cutoff for the smooth cost. We use $L = 10$ for the upper-bound of depth discontinuity in a label space with 100 different labels. Here, the weight function $\omega(p, q, \lambda)$ is slightly different since we use a weight factor λ to balance the spatial proximity and color similarity. Most importantly, λ is a variable in the iteration.

$$\omega(p, q, \lambda) = e^{-\|p-q\|_2} + \lambda e^{-\|\hat{I}_{ref}(p) - \hat{I}_{ref}(q)\|_2} \tag{9}$$

Finally, our energy function of depth reconstruction can be summarized as

$$\begin{aligned} E &= E_{d_data} + E_{d_cons} \\ &= \sum_{p \in I_{ref}} F(p, d) + \sum_{\{p,q\} \in \mathcal{N}} V_{\{p,q\}}(d(p), d(q)) \end{aligned} \tag{10}$$

Reconstructing scene depth D and appearance I_{ref} is a typical "chicken-and-egg" problem. Thus, we need to find out an approximate but efficient way to solve (4). To avoid explicitly solving $B_k(O, d)$, we obtain the matching cost on a specific cluster of candidates C_s, which makes a fair chance to include non-occluded pixels as many as possible. The relationship between (4) and (10) is that, we use in-class distance dt_s as an approximation of $\sum_{k=1}^{K-n}\|H_k^{-1}(d)\hat{I}_k - \hat{I}_{ref}\|_2$, at the cost of $\rho(n) = \rho(K-n) = 1/ct_s$. This approximation is reasonable in most of cases, depending on that, at the correct depth, the pixels of foreground occlusions are randomly distributed, whereas non-occluded pixels of the targets are highly clustered, as shown the case at depth d_0 in Fig. 6.

So far, we have discussed on scene depth recovery under foreground occlusions. Please notice that, the occlusion-removed image of the reference view $\hat{I}_{ref}$ is not known at this stage. We are forced to use the clustering centroid of C_s as the reasonable approximation of $\hat{I}_{ref}$, here we denote as $\hat{I}^t_{ref}$, s represents the round of iterations. Then, we can apply a pixel-wise selecting $\hat{I}^t_{ref}(d_{i,j})$ according to the above initialized depth estimation. For simplicity, we can directly fill the pixel $\hat{I}^t_{ref}(d_{i,j})$ by using the corresponding pixel of synthetic aperture photography sequence, which is described in (5). But, this straight forward $\hat{I}^t_{ref}(d_{i,j})$ selecting process may be deteriorated by inaccurate depth estimation pixels, which are general existance in most of 3D reconstructions. For this reason, we also introduce a global optimization solution to obtain better $\hat{I}_{ref}$ in an iterative framework. Similar to (10), we propose the energy function of appearance reconstruction as below,

$$\begin{aligned} E &= E_{a_data} + E_{a_cons} \\ &= \sum_{p \in I_{ref}} R(p, l) + \sum_{\{p,q\} \in \mathcal{N}} U_{\{p,q\}}(l(p), l(q)) \end{aligned} \tag{11}$$

where $R(p, l)$ represents the appearance reconstruction data cost of point p, and $U_{\{p,q\}}(l(p), l(q))$ is corresponding smooth cost. The data cost $R(p, l)$ is calculated from the color difference between $\hat{I}^t_{ref}(p)$ and the possible candidate color intensity grade l as below.

$$R(p, l) = min(\|l - \hat{I}^t_{ref}(p)\|_2, T_c) \tag{12}$$

where T_c is a cut-off threshold, and $T_c = 100$ through our experiments. Unlike the smooth term in depth estimation, we use a Gaussion weight based smooth cost in appearance estimation, which is similar as the weight function of adaptive supported window in [52].

$$U_{p,q}(l(p), l(q)) = 12.5 \times e^{-\frac{(p-q)^2}{2.0}} \times \min(L_c, |l(p) - l(q)|) \tag{13}$$

where we set $L_c = 100$ for $[0, 255]$ intensity level of $8bit$ RGB images. The attentive reader may notice that the energy minimization results tend to close all-in-focus synthetic aperture photography by using the data cost of (12). But, the image contrast of SAP will be greatly decreased by the low frequency energy of defocused foreground occlusions. By using the integration process of (5), the foreground energy will be scattered all over the $\hat{I}^t_{ref}$. The traditional SAP process includes two major steps, which are summing and then averaging on sample candidates, please see in (5). As intuitively shown in Fig. 5, all angular samples contribute to the SAP result, including both occluded and non-occluded ones, so that the pixels of foreground will be averaged into SAP pixels of background targets,

Algorithm 1: Iterative Reconstructing Algorithm for Partially Occluded Scene.

Require: source image set I, consistency threshold T, depth range $[d_a, d_b]$, the number of K-means clusters m, the max iterative number t_{max}, weight factor λ

Ensure: depth map d^* and scene appearance I^*_{ref}

for each $d \in [d_a, d_b]$, and each pixel p **do**
  calculate data cost $F(p, d)$ by (6);
end for
optimize d^0 using graph-cuts without color similarity term, as in [12];
estimate $\hat{I}^0_{ref} = \frac{1}{K}\sum_{k=1}^{K} H_k^{-1}(d^0) I_k$, by using (5);
while $t < t_{max}$ **do**
  for each $d \in [d_a, d_b]$, and each $(p, q) \in \mathcal{N}$ **do**
    calculate $V_{p,q}(d(p), d(q))$ from $\hat{I}^{t-1}_{ref}$ by (8);
  end for
  optimize d^t by minimizing (10);
  for each color channel **do**
    for each $l \in [0, 255]$, and each $(p, q) \in \mathcal{N}$ **do**
      calculate $R(p, l)$ and $U_{p,q}(l(p), l(q))$ from $\hat{I}^{t-1}_{ref}$ by (13) and (14);
    end for
  end for
  optimize $\hat{I}^t_{ref}$ by minimizing (11);
  $\lambda \leftarrow \lambda * (0.7)^{t-1}$;
  $t \leftarrow t + 1$;
end while
$d^* \leftarrow d^t, I^*_{ref} \leftarrow \hat{I}^t_{ref}$

which leads low contrast result. More details about SAP process can be found in [28] and [31], which conduct light field imaging on spatial-domain and transform-domain respectively. As shown in Fig. 1(c), the target scene looks like to be seen through dark glasses when using traditional SAP algorithm. In order to suppress the unexpected dark-glasses-artifact, we can use the clustering centroid instead of directly using $\hat{I}^t_{ref}$. In this case, for a point p, its appearance reconstruction data cost can be rewritten as

$$R(p, l) = min(\|l - \mu_i\|_2, T_c) i \in (1, 2, \ldots, m) \tag{14}$$

where μ_i represents a possible candidate color from the i_{th} order clustering centroid.

Now we summarize our method in Algorithm 1. At first, $\hat{I}_{ref}$ is not available, so that we force $\lambda = 0$ and consider only the geometry smoothness constraint when calculate $V_{p,q}((d(p), d(q))$. After we obtain the first estimation of $\hat{I}^0_{ref}$, the color smoothness can be employed as an constraint term in occlusion handling for two reasons. First, the occlusion can be blurred out by imaging with a camera array [45], so that the color appearance is not very sensitive to occlusions. Second, if the depth estimation is inaccurate, slight defocus blur is acceptable in the color appearance estimation. Thus, we use this color based constraint to smooth the depth estimation. One side-effect is the result would be over-smoothed. Therefore, we use a coarse-to-fine strategy to recover depth by decreasing λ gradually in the iteration.

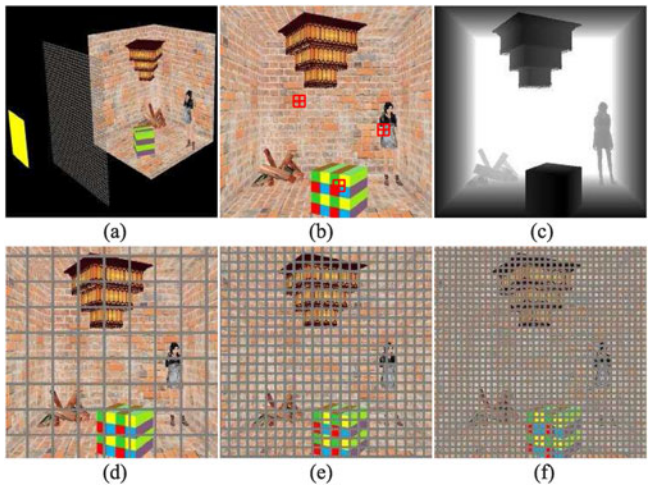


Fig. 7. Synthetic scene with foreground occlusions. (a) The cave scene. (b) One sample of 64 views. (c) Ground truth of depth. (d)-(f) Cave scenes with 20%, 50%, 70% occlusions.

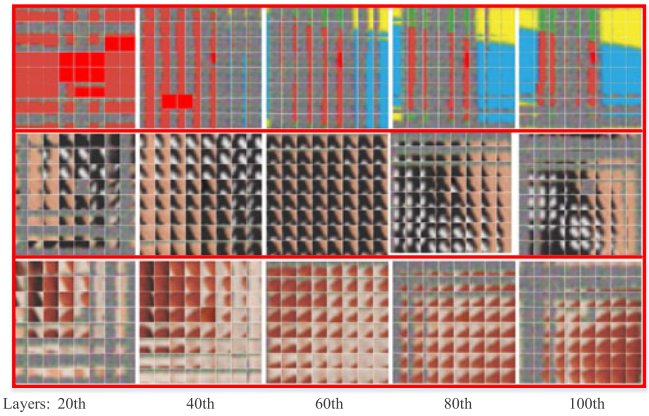


Fig. 8. The variation of 64 angular samples on different layers by using synthetic OpenGL data.

V. Experimental Results

In order to compare with existing methods quantitatively, we first conduct experiments on the synthetic camera array data. We render a Brickwall room using OpenGL, then the scene is captured with an 8×8 equidistant synthetic camera array. For each camera, we generate a 512×512 resolution image with 45° field of view. As in [45], the foreground occluder is composed of horizontal and vertical bars, which can be precisely controlled by varying the adjacent spacing between the bars. We draw the foreground occlusion bars with different noise (white noise, pink noise and uniform noise), and controlled the amount of occlusion from 10% to 70%, as shown in Fig. 7.

To intuitively illustrate foreground occlusion artifacts, we plot the variation of angular samples with same refocused position on different depth layers, as shown in Fig. 8. Each patch shows a comparison of samples from 64 views, the locations of original patches are marked in red, as shown in Fig. 7(b). From Fig. 8, we can see that the different angular samples will appear a perfect similarity on the correct depth layer, e.g., the central patch in middle row. Due to foreground occlusion, the photo consistency of angular samples will be hardly hold among different angular

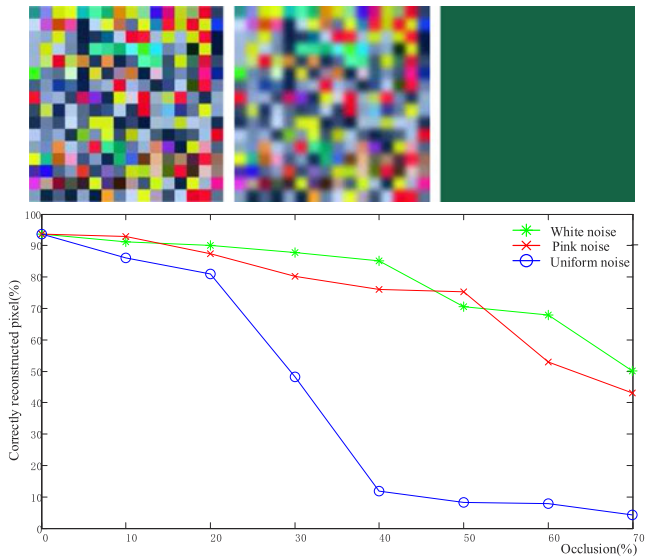


Fig. 9. Comparison of depth reconstruction under white, pink, uniform noisy occlusion.

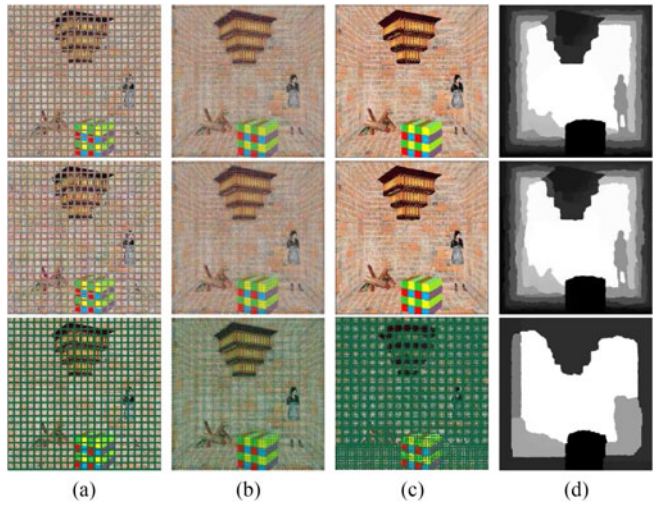


Fig. 10. Scene reconstruction with different foreground occlusion patterns. (a) 40% different foreground occlusion patterns. (b) shows scene appearance reconstruction with data cost in (12). (c) shows scene appearance reconstruction with data cost in (14). (d) shows scene depth reconstruction.

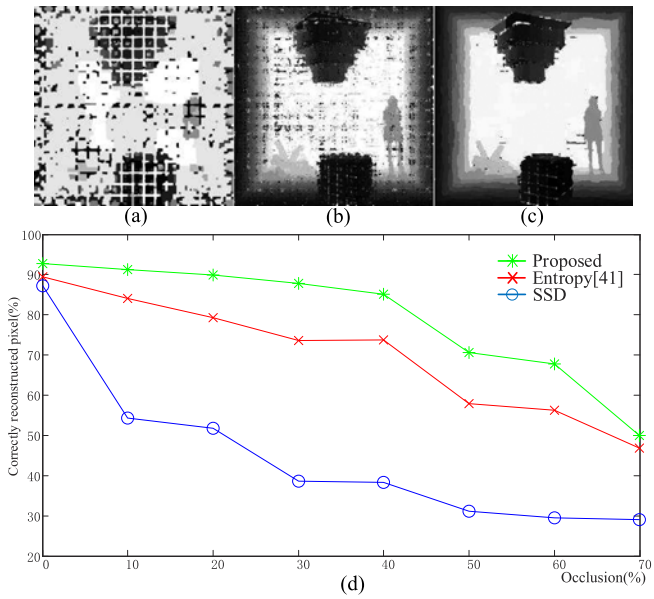


Fig. 11. Depth recovery on synthetic data. (a)–(c) Depth map by SSD, entropy based approach and ours under 40% occlusion. (d) The percentage of correctly reconstructed pixel of three approaches under different occlusion proportions.

samples, even on its correct depth layer, as the 20th layer patch in the top row and the 80th layer patch in the bottom row. Fortunately, the photo consistency still exists in most parts of non-occluded angular samples.

To simulate the complexity of real world environment, we test the proposed algorithm under three different types of noise occlusions, as shown in Fig. 9. On the synthetic OpenGL scene data, we calculate the percentages of correctly depth reconstructed pixels under occlusion, in which the occlusion rates is increased from 10% to 70% at equal interval. Here, we divide the scene depth space into 100 different layers. Then, we set our error tolerance limit to 5 depth layers aparting from ground truth, which means 5% error is acceptable within scene depth space. From left to right, the top row of Fig. 9 shows the patterns of white noise, pink noise and uniform colored noise respectively. We can see that our algorithm is well performed under white noise and pink noise occlusions, but the reconstruction accuracy drops fast when the occlusion proportion is higher than 50%. In the uniform colored occlusion case, failure point of the proposed algorithm would start at 20% occlusion. That because uniform colored occlusion appears highly photo consistency in the K-means process, which may be falsely recognized as a valid target point. Fortunately, since the nature lights is usually randomly illuminating and shading on the geometrical surfaces of 3D occlusions, the uniform colored occlusion is much rare than non-uniform colored cases in nature environment.

To further explain the characteristics of proposed algorithm, we show the scene reconstruction results at 40% white noise, pink noise and uniform colored noise occlusions in Fig. 10, from which we can see that our algorithm is robust to both white noise and pink noise occlusions. But as we mentioned above, the high similarity of foreground occlusion will lead two highly clustered centers, so that our algorithm can be vulnerable with uniformed colored noises, as shown the bottom row in Fig. 10. The Fig. 10(b) and (c) show the appearance reconstruction, the brighter results in Fig. 10(c) are using the cost in (14), in which the dark-glasses-artifact is significantly suppressed than that in Fig. 10(b). The Fig. 10(d) is corresponding depth recovery results.

In Fig. 11, we compare the proposed approach with classical sum-squared difference (SSD) approach and entropy based approach [45]. Figs. 11(a)–(c) show the depth recovery results under 40% occlusion and all results are optimized using graph-cuts [14]. To evaluate matching costs independently, we conduct a comparison using winner-take-all strategy rather than global optimization in Fig. 11(d). We can see that the K-means based matching cost obviously outperforms other two matching costs.

In Fig. 12, we demonstrate our method using a real planar camera array. The elemental CCD camera (CK-IH046C) has a 752×576 resolution and 37.0° field of view with a 12 mm

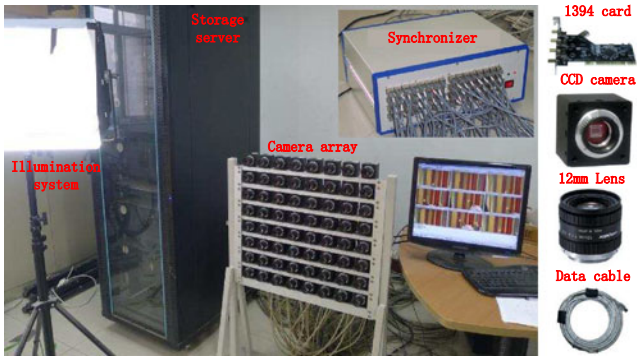


Fig. 12. Real 8×8 camera array.

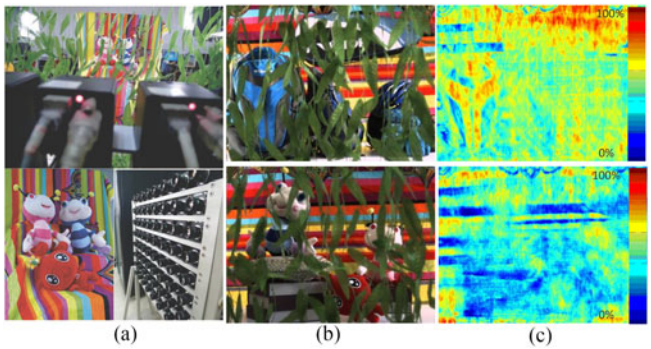


Fig. 13. Target scene under moderate severe foreground occlusions. (a) The setup of the experiment. (b) The central camera views of two datasets. (c) The corresponding proportion map of occlusion.

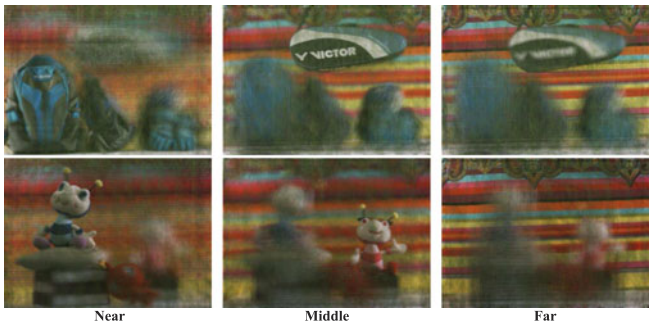


Fig. 14. Traditional synthetic aperture photography on different focal planes.

lens. The baseline between two adjacent cameras is 70 mm, as shown in Fig. 12. The camera array is triggered by numerical controlled synchronizer. In Fig. 13(a), we show the experimental setup under moderate severe foreground occlusions, the central camera views of target scene and its corresponding occlusion ratio estimation are shown in Fig. 13(b) and (c) respectively. The traditional SAP results on three different focal planes are shown in Fig. 14, from which we can see the foreground occlusion can be generally blurred out, and only those targets on a specific plane are sharp focused. We only give out three out of one hundred SAP images from near focus to far background.

In Fig. 15, we compare our iterative scene appearance reconstruction with existing methods. We show two groups of our scene appearance reconstruction using the (12) and (14) cost function in Fig. 15(d) and (e) respectively. Our scene appearance reconstruction can generate more clearer results than entropy based methods in [45]. Another advantage of proposed appearance reconstruction is that we can produce the all-in-focus scene appearance, which is not realized in Pei's work in [48]. Although the occlusion removal via pixel labeling can effective reduce the dark-glasses-artifact in [48], some foreground energy still be unavoidable distributed to the results in Fig. 15(c). This is because the pixel labeling process may failure while the foreground occlusion texture exists highly color similarity. Even when pixel labeling is well performed, another side-effect can not be ignored, the traditional integration in [48] may greatly deteriorate appearance reconstruction as introducing strong aliasing artifacts. Please see the out-of-focus region in Fig. 15(c), the aliasing is led by angular down-sampling, the interested reader is referred to paper [53] for detailed information. We give out a 3-channels RGB image histogram comparison on the bottom of Fig. 15. These histograms show that, judging from the pixels tonal distribution, our scene appearance recovery using (14) are more close to the non-occluded ground truth. especially on the green color channel, which be significantly entangled by foreground occlusion. To analyze on the image structure recovery performance, we have further employed the structure similarity (SSIM) index [54] as a metric of image appearance reconstruction. With the results in Fig. 15, we plot the structural differences between results of appearance reconstruction and the non-occluded ground truth in Fig. 16. The white region of Fig. 15 represents high structure similarity, whereas the dark region represents failure appearance reconstruction. Apparently, the proposed reconstruction preserves more scene details than other algorithms. In general, compare to the work in [48], our algorithm can produce similar but more bright appearance reconstruction without foreground modeling, which is still a challenge problem in most of practical applications.

Because we have not obtained the depth ground truth, in the experiment of depth recovery under moderate severe foreground occlusion, we only compare the results between our approach and the approach in [45]. Since the global optimization was not employed in [45], we improved their approach using graph-cuts, and made the results more comparable. Taking the depth estimation from non-occluded data as benchmark in Fig. 17(a), our approach generates better scene depth (Fig. 17(d)) than traditional multiview occlusion handling [14] and entropy based method [45], as shown in Fig. 17(b) and (c). From Fig. 17 (b) we can see that, even preventing the violation of visibility constraint on depth assignment, the depth reconstruction will be still badly failed because none of views can be treated as the reference. The results of entropy based method will also become noisy when the foreground consisting of similar colored occlusions, as in Fig. 17(c). Besides, we have tested some light field scene reconstruction algorithms, but none of them can obtain satisfactory results by applying the release codes in [36], [40]. The light field algorithms are more suitable for compact light field camera data, in which the disparities is nearly in linear distribution, thus the light field EPI images are consisted of lines with different slopes depending on depth. By accurate calibrating of different angular elemental cameras, our algorithm can handle nonlinear and non-uniform disparities distribution, which is different from microlens array plenoptic cameras cases. For this reason,

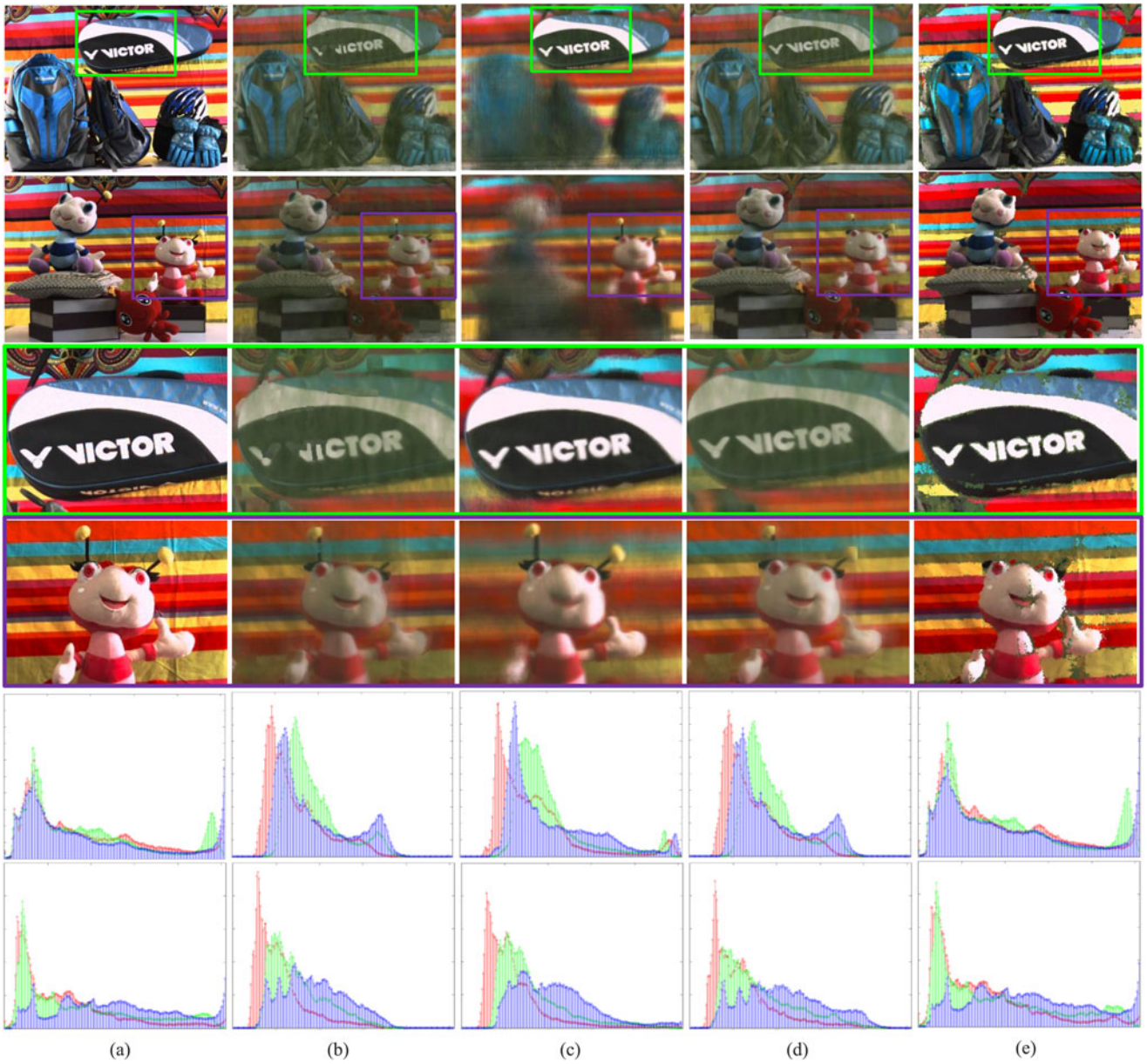


Fig. 15. Scene appearance recovery results on real data. (a) The non-occluded ground truth. (b) Entropy based scene appearance recovery using method in [45]. (c) Pixel labeling scene appearance recovery in [48]. (d) Our iterative scene appearance recovery using (12). (e) Our scene appearance recovery using (14).

the major concern of popular light field algorithms are handling self occlusions on objects boundaries, but we focus on handling foreground occlusion.

In our experiments, the selection of weight factor λ is scene dependent, which can be optimized as suggested in [55]. At the same time, we set photo consistency upper bound $T = 200$. The threshold T is chosen empirically large because we have to compromise with large angular variations when using camera array data. In our 8×8 camera array, we set the number of clusters $m = 5$ in K-means, which is not only depended on the number of views, but also relied on the texture complexity of foreground occlusions. Intuitively, we need a bigger m when the number of views is large or the foreground occlusions is strong textured. In order to choose a suitable m, we analyse the relationship between reconstruction accuracy and m value using synthetic dataset in Fig. 18. Apparently, we observe that a bigger m will basically lead a more accurate result, but a large m will cost much more runtime, e.g., $m = 20$ costs 10 times more runtime to obtain a comparable result with $m = 5$. Notice that, the overfitting of clustering may occur when m is pretty close to the number of angular samples, we also show an overfitting example case in Fig. 19. As shown in the last subimage of Fig. 19, the number of clustering centers $m = 60$ is pretty close to the number of camera 64, the reconstruction accuracy is decreased by the overfitting of clustering.

Furthermore, we show iterative reconstruction results of camera array data in Fig. 20. Even when the depth map is not accurately estimated, the reconstructed appearance is still acceptable as we mentioned above. In the iteration process of Algorithm 1, we choose $t_{max} = 3$ since there is less contributions as increas-

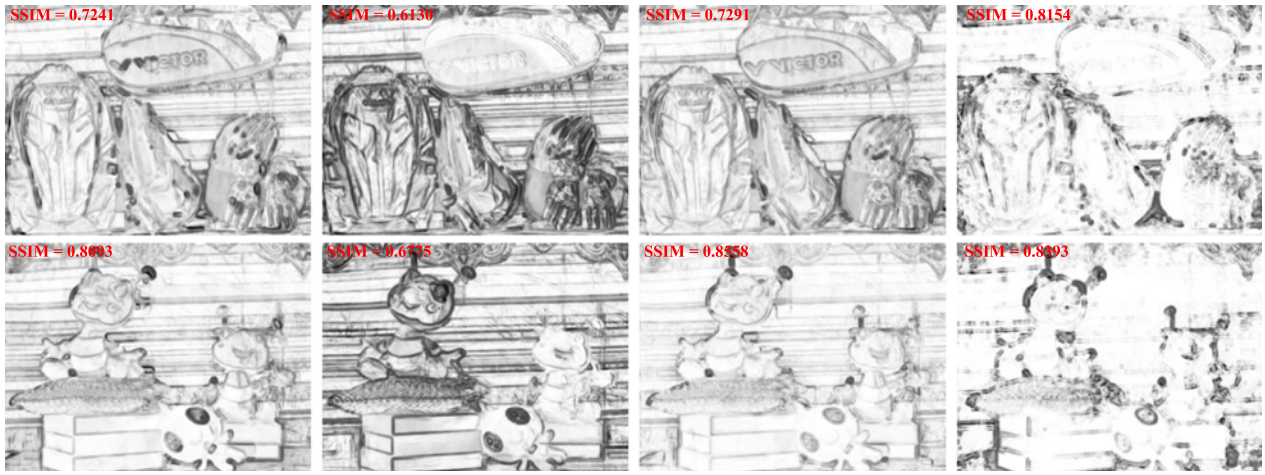


Fig. 16. Structural similarity comparison between different appearance recovery results and non-occluded ground truth. (a) Structural similarity between Fig. 15(b) and (a). (b) Structural similarity between Fig. 15(c) and (a). (d) Structural similarity between Fig. 15(d) and (a). (e) Structural similarity between Fig. 15(e) and (a).

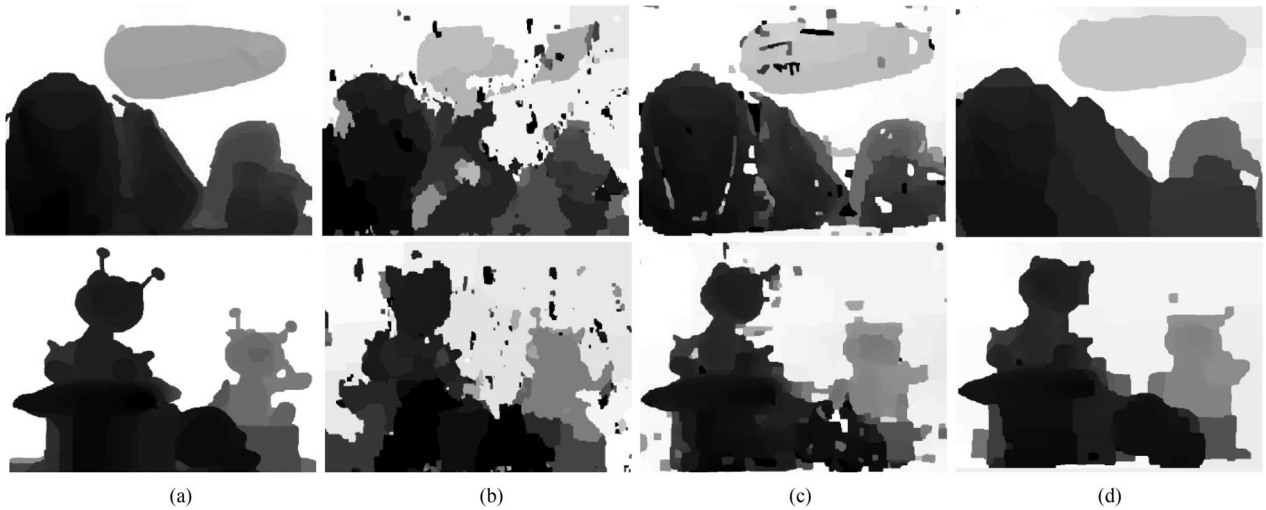


Fig. 17. Scene depth recovery results on real data. (a) Estimated depth map from occlusion free dataset. (b) Recovered depth map using visibility constraint in multiview occlusion handling [14]. (c) Recovered depth map using entropy based method [45]. (d) Recovered depth map using our method.

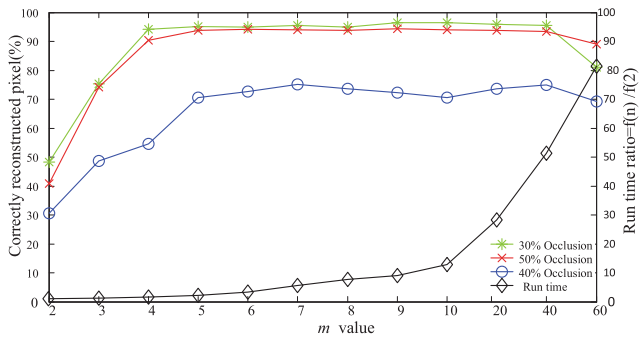


Fig. 18. The analysis on depth reconstruction accuracy and runtime with different m value.

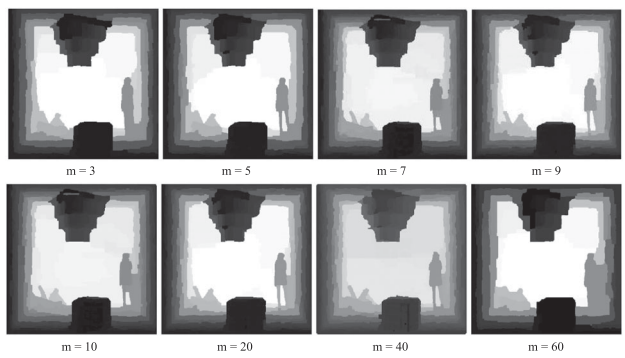


Fig. 19. Depth reconstruction results on synthetic cave scenes with different m value.

ing t_{max}. In Fig. 20, both the scene depth and appearance are reconstructed by using our coarse-to-fine iteration strategy. The top row scene appearance reconstruction of Fig. 20 is computed using data cost in (12), thus the energy of foreground occlusion is spread all over the images, which leads the above mentioned dark-glasses-artifact. We employ the data cost in (14) again to reduce these artifacts, as shown in the middle row of Fig. 20. In Table I, we give out a runtime comparison on four different

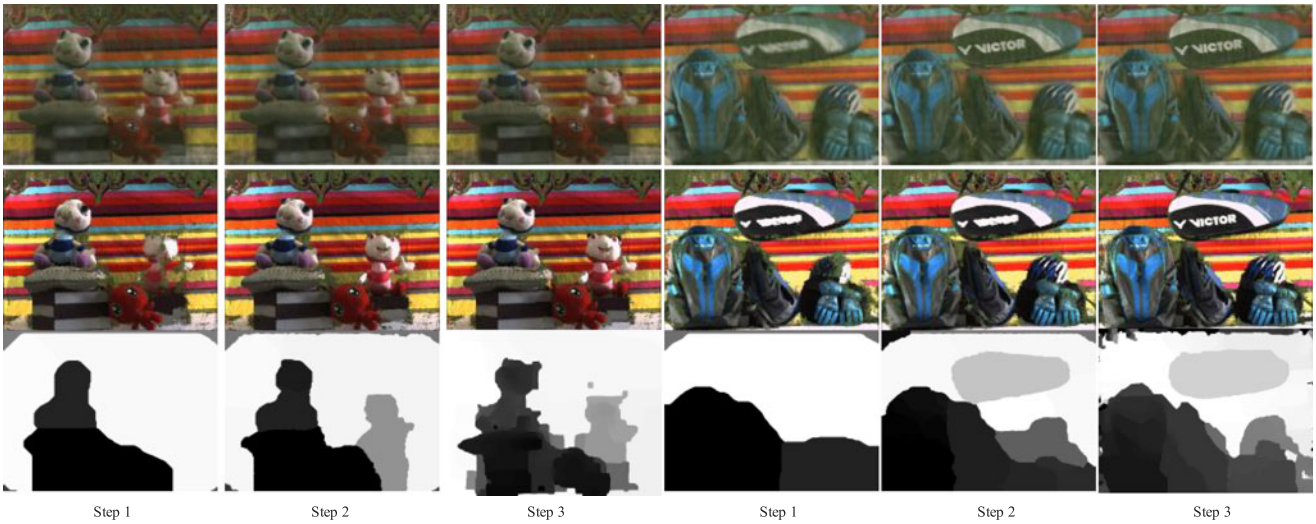


Fig. 20. The results of 3 steps iteration.

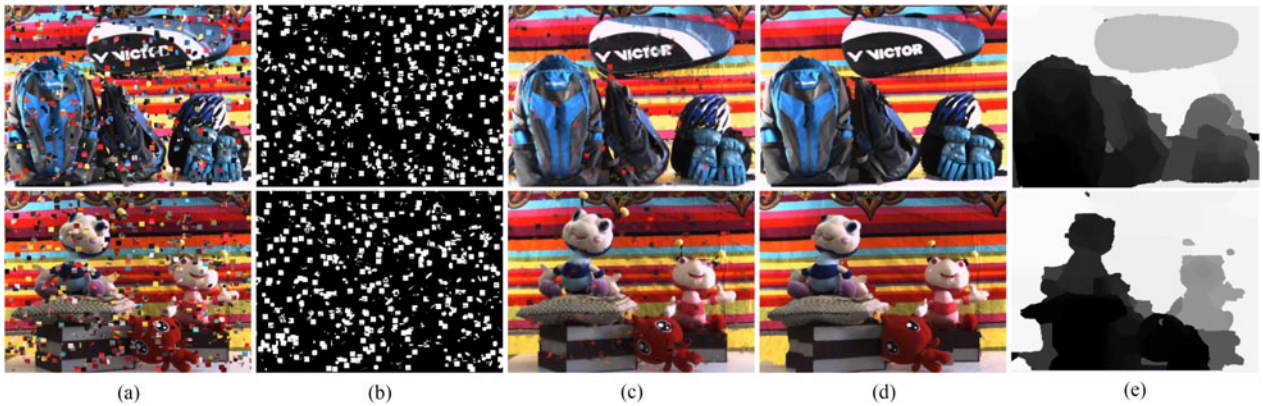


Fig. 21. Scene reconstruction results on real data with raindrops like noises. (a) Adding noises to non-occluded scene in Fig. 15(a). (b) The mask image of noises. (c) Patch based image inpainting results using algorithm in [24]. (d) Recovered scene appearance using our method. (e) Recovered depth map using our method.

TABLE I
RUNTIME COMPARISON ON FOUR DIFFERENT METHODS

Runtime	Traditional SSD	Entropy [45]	Pixel labeling [48]	Proposed
Cave $512 * 512$	1.00	0.31	112.04	**13.20**
Bags $752 * 576$	1.72	0.49	188.52	**24.29**
Toys $752 * 576$	1.70	0.43	171.16	**22.85**

algorithms. Taking the runtime of most common sum-squared difference (SSD) algorithm as a baseline, we calculate the runtime ratio between other algorithms and classic SSD. Since our scene reconstruction is conducted on large number of views, so entropy based algorithm is even faster than SSD algorithm. Without foreground occlusion modeling, the proposed algorithm is much more efficient than foreground occlusion pixel labeling algorithm.

As we see in Fig. 13, under such dense green branches occlusion, scene reconstruction is still an open and challenging task. Compare to synthetic scene case, the very approximate color and depth of green branches will lead worse reconstruction results. We are inspired by the work of raindrops removal in [56], in which the raindrops can be deemed as a classic kind of foreground occlusion with random distribution. So we tent to apply the proposed algorithm for this raindrops removal case. Unfortunately, it is not feasible to move our large and heavy camera array outside at this stage. But, we have added some fictitious raindrop like noises to real scene data, as shown in Fig. 21(a). We show image inpainting results using algorithm in [24] and the proposed reconstruction in Fig. 21(c)–(d) respectively. Please see in Fig. 21(d), the proposed depth reconstruction can also obtain acceptable results under the raindrops like foreground occlusion. In most of practical applications, it is hardly to get the occlusion mask prior, which is not needed in our algorithm. Moreover, even the occlusion mask prior is obtainable, the image inpainting techniques may also fail to reconstruct heavily occluded scene. For example, we have tried both pixel-wised and patch-wised single image inpainting algorithms [24], [57] to recovery our real scene in Fig. 13, but none of results is satisfactory, as shown in Fig. 22.

Attentive readers will also find an unexpected side-effect with this dark-glasses-artifact reduction using (14), the appearance reconstruction may randomly generate false results as encoun-

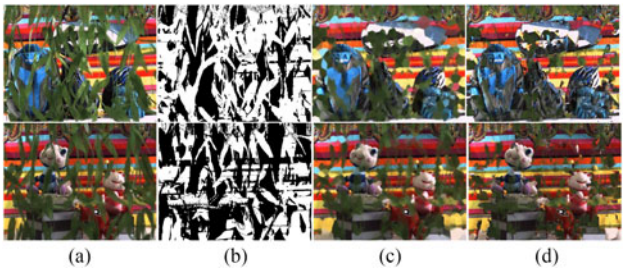


Fig. 22. Image inpaiting results on real data. (a) The target images. (b) The mask image of (a). (c) Pixel-wise image inpainting [57]. (d) Patch based image inpainting [24].

tering multiple layers of occlusions. For example, we can find that the antennas of red toy ants disappear in the appearance recovery, as shown in Fig. 15(e). From our analysis, the tiny antennas of toy ants may be eliminated as occlusions relative to the background fabric. Thus, if there are lots of multiple layers of occlusions in the target scene, and the accurate scene appearances reconstruction is the major concern, we recommend giving up data cost in (14) to avoid this false positive foreground occlusion removal. Fortunately, the data cost in (12) will still work correctly as other existing methods, and our algorithm can produce both accurate scene depth and all-in-focus appearances without occlusion modeling.

VI. Conclusion

In this paper, we bring out a new scene reconstruction problem, in which the REC assumption is untenable. To deal with severe foreground occlusion, we propose a generalized scene reconstruction model for recovering both scene depth and appearance. Since none of source images can be used as the reference view, we apply the synthetic aperture photography to estimate a reasonable reference with an inaccurate depth prior. The proposed scene reconstruction model allow us to minimize the contributions of the unexpected occluded views.

Further, we introduce an iterative scene reconstruction algorithm, which is more robust to foreground occlusion than existing algorithms. Most importantly, our approach is free of occlusion modeling and detecting. We introduce a K-means clustering to separate target scene from heavy foreground occlusion. Therefore, we need more views to ensure the validity of proposed approach. The proposed algorithm performs well under white noise and pink noise occlusion, but reconstruction accuracy drops quickly when the proportion of occlusion is above 50%.

In the experiment part, we verify the proposed algorithm on both synthetic and real datasets. Then, we analyse the selection of parameters, such as the count of K-means centers, iteration times, photo consistency upper bounds, etc. Most importantly, the results show that our algorithm can successfully recover accurate scene depth, as well as occlusion removed scene appearance. Compare to existing methods, the quantitatively analysis indicates the robustness and effectiveness of the proposed algorithm. In future, we will research on parameters adaptive selecting, and apply our approach on object tracking and detection in presence of foreground occlusion.

References

[1] M. L. Mata and D. L. Ringach, "Spatial overlap of on and off subregions and its relation to response modulation ratio in macaque primary visual cortex," *J. Neurophysiol.*, vol. 93, no. 2, pp. 919–928, 2005.
[2] S. M. Shuwairi, C. E. Curtis, and S. P. Johnson, "Neural substrates of dynamic object occlusion," *J. Cogn. Neurosci.*, vol. 19, no. 8, pp. 1275–1285, 2007.
[3] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 7–42, 2002.
[4] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2006, pp. 519–528.
[5] D. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science*, 194(4262), pp. 283–287, 1976.
[6] C. L. Zitnick and T. Kanade, "A cooperative algorithm for stereo matching and occlusion detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 7, pp. 675–684, Jul. 2000.
[7] D. Geiger, B. Ladendorf, and A. Yuille, "Occlusions and binocular stereo," in *Proc. Eur. Conf. Comput. Vis.*, 1992, pp. 425–433.
[8] J. Sun, Y. Li, S. Kang, and H. Shum, "Symmetric stereo matching for occlusion handling," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2005, pp. 399–406.
[9] M. Brown, D. Burschka, and G. Hager, "Advances in computational stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 993–1008, Aug. 2003.
[10] M. Bleyer and M. Gelautz, "A layered stereo algorithm using image segmentation and global visibility constraints," in *Proc. Int. Conf. Image Process.*, 2004, pp. 2997–3000.
[11] D. Min and K. Sohn, "Cost aggregation and occlusion handling with WLS in stereo matching," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1431–1442, Aug. 2008.
[12] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Proc. Int. Conf. Comput. Vis.*, 2001, pp. 508–515.
[13] G. Vogiatzis, C. Hernández Esteban, P. Torr, and R. Cipolla, "Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2241–2246, Dec. 2007.
[14] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 82–96.
[15] J. Sun, N. Zheng, and H. Shum, "Stereo matching using belief propagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 787–800, Jul. 2003.
[16] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Proc. Int. Conf. Pattern Recognit.*, 2006, pp. 15–18.
[17] Q. Yang, L. Wang, R. Yang, H. Stewénius, and D. Nistér, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 492–504, Mar. 2009.
[18] A. F. Bobick and S. S. Intille, "Large occlusion stereo," *Int. J. Comput. Vis.*, vol. 33, no. 3, pp. 181–200, 1999.
[19] J. C. Kim, K. M. Lee, B. T. Choi, and S. U. Lee, "A dense stereo matching using two-pass dynamic programming with generalized ground control points," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 1075–1082.
[20] A. Criminisi, A. Blake, C. Rother, J. Shotton, and P. H. Torr, "Efficient dense stereo with occlusions for new view-synthesis by four-state dynamic programming," *Int. J. Comput. Vis.*, vol. 71, no. 1, pp. 89–110, 2007.
[21] S. Huq, A. Koschan, and M. Abidi, "Occlusion filling in stereo: Theory and experiments," *Comput. Vis. Image Understanding*, vol. 117, no. 6, pp. 688–704, 2013.
[22] S. Krinidis and V. Chatzis, "Fuzzy energy-based active contours," *IEEE Trans. Image Process.*, vol. 18, no. 12, pp. 2747–2755, Dec. 2009.
[23] H. Jin, Y. Li, B. Xing, and L. Wang, "A geometric image segmentation method based on a bi-convex, fuzzy, variational principle with teaching-learning optimization," *J. Intell. Fuzzy Syst.*, vol. 31, no. 6, pp. 3075–3081, 2016.
[24] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[25] T. Ruzic and A. Pizurica, "Context-aware patch-based image inpainting using Markov random field modeling," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 444–456, Jan. 2015.
[26] E. H. Adelson and J. R. Bergen, *The Plenoptic Function and the Elements of Early Vision*. Cambridge, MA, USA: MIT Press, 1991.
[27] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. Annu. Conf. Comput. Graph. Interactive Techn.*, 1996, pp. 31–42.
[28] B. Wilburn *et al.*, "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, 2005.
[29] Lytro, 2016. [Online]. Available: http://www.lytro.com
[30] Raytrix, 2016. [Online]. Available: http://www.raytrix.com
[31] R. Ng, "Fourier slice photography," in *Proc. ACM SIGGRAPH*, 2005, pp. 735–744.
[32] A. Levin, S. W. Hasinoff, P. Green, F. Durand, and W. T. Freeman, "4D frequency analysis of computational cameras for depth of field extension," *ACM Trans. Graph.*, vol. 28, pp. 1–14, 2009.
[33] T. E. Bishop and P. Favaro, "Full-resolution depth map estimation from an aliased plenoptic light field," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 186–200.
[34] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 606–619, Mar. 2014.
[35] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, 2013, Art. no. 73.
[36] M. W. Tao, P. P. Srinivasan, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi, "Depth from shading, defocus, and correspondence using light-field angular coherence," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1940–1948.
[37] H. G. Jeon *et al.*, "Accurate depth map estimation from a lenslet light field camera," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1547–1555.
[38] H. Lin, C. Chen, S. B. Kang, and J. Yu, "Depth recovery from light field using focal stack symmetry," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 3451–3459.
[39] C. Chen, H. Lin, Z. Yu, S. B. Kang, and J. Yu, "Light field stereo matching using bilateral statistics of surface cameras," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1518–1525.
[40] T. C. Wang, A. A. Efros, and R. Ramamoorthi, "Depth estimation with occlusion modeling using light-field cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2170–2181, Nov. 2016.
[41] Williem and I. K. Park, "Robust light field depth estimation for noisy scene with occlusion," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4396–4404.
[42] L. Diaz-Mas, R. MuOz-Salinas, F. J. Madrid-Cuevas, and R. Medina-Carnicer, "Shape from silhouette using Dempster–Shafer theory," *Pattern Recognit.*, vol. 43, no. 6, pp. 2119–2131, 2010.
[43] X. Wang, J. Shi, H. S. Park, and Q. Wang, "Motion-based temporal alignment of independently moving cameras," *IEEE Trans. Circuits Syst. Video Technol.*, vol. PP, no. 99, pp. 1–1, doi:10.1109/TCSVT.2016.2581659.
[44] M. Levoy, B. Chen, V. Vaish, M. Horowitz, I. Mcdowall, and M. Bolas, "Synthetic aperture confocal imaging," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 825–834, 2004.
[45] V. Vaish, M. Levoy, R. Szeliski, C. L. Zitnick, and S. B. Kang, "Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 2331–2338.
[46] C. M. Do and B. Javidi, "3D integral imaging reconstruction of occluded objects using independent component analysis-based k-means clustering," *J. Display Technol.*, vol. 6, no. 7, pp. 257–262, 2010.
[47] S. Mccloskey, "Masking light fields to remove partial occlusion," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 2053–2058.
[48] Z. Pei, Y. Zhang, X. Chen, and Y. H. Yang, "Synthetic aperture imaging using pixel labeling via energy minimization," *Pattern Recognit.*, vol. 46, no. 1, pp. 174–187, 2013.
[49] H. Zhu, Q. Wang, and J. Yu, "Occlusion-model guided anti-occlusion depth estimation in light field," *IEEE J. Selected Topics Signal Process.*, 2017.
[50] C. R. Dyer and S. M. Seitz, "Photorealistic scene reconstruction by voxel coloring," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 1997, pp. 1067–1073.
[51] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Appl. Statist.*, vol. 28, pp. 100–108, 1979.
[52] K. J. Yoon and I. S. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, Apr. 2006.
[53] Z. Xiao, Q. Wang, G. Zhou, and J. Yu, "Aliasing detection and reduction in plenoptic imaging," in *Proc. IEEE Conf. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3326–3333.
[54] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[55] L. Zhang and S. Seitz, "Parameter estimation for MRF stereo," in *Proc. Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, 2005, pp. 288–295.
[56] S. You, R. T. Tan, R. Kawakami, and K. Ikeuchi, "Adherent raindrop detection and removal in video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1721–1733, Sep. 2016.
[57] A. Telea, "An image inpainting technique based on the fast marching method," *J. Graph. Tools*, vol. 9, no. 1, pp. 23–34, 2004.



Zhaolin Xiao received the Bachelor's and Master's degrees from the Department of Computer Science and Engineering, Xi'an Technological University, Xi'an, China, in 2006 and 2009, respectively, and the Ph.D. degree from the Department of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China, in 2014. He then joined Xi'an University of Technology as a Lecturer. He is currently a member of China Computer Federation. In 2011 and 2012, he worked as a visiting scholar in the School of Computer Science, The University of Wisconsin-Madison, and the Department of Computer Science, University of Delaware, for one month. His research interests include image processing, computer vision, and computational photography, and his current research also focuses on image-based 3-D reconstruction and light field imaging and processing.



Lipeng Si received the Bachelor's and Master's degrees from the Department of Engineering Physics, TsingHua University, Beijing, China, in 2004 and 2007, respectively. He is currently a Ph.D. candidate in the Department of Computer Science and Engineering, Northwestern Polytechnical University. His research interests include computer vision, image processing, and computational photography, and his current research also focuses on image-based 3-D reconstruction and light field imaging and processing.



Guoqing Zhou received the B.S. degree in computer science and Master's and Ph.D. degrees from the School of Computer Science, Northwestern Polytechnical University, Xi'an, China, in 2003, 2009, and 2013. He joined Northwestern Polytechnical University as a Lecturer. He is currently an Associate Professor in the School of Computer Science, Northwestern Polytechnical University. His research interests include computer vision and computational photography, such as 3-D reconstruction, global optimization, and light field imaging and processing. He has published more than ten papers in the international journals and conferences, such as ICCV, CVPR.

Single Aperture Spectral+ToF Compressive Camera: Toward Hyperspectral+Depth Imagery

Hoover Rueda, *Student Member, IEEE*, Chen Fu, *Student Member, IEEE*, Daniel L. Lau, *Senior Member, IEEE*, and Gonzalo R. Arce, *Fellow, IEEE*

***Abstract*—Spectral imaging involves the sensing of a large amount of spatial information across a multitude of wavelengths. Conventional approaches rely on scanning techniques to construct a spectral data cube. Recently, compressive spectral imaging (CSI) has allowed to estimate spectral images with as few as a single coded snapshot. On a different front, 3-D ranging imaging often involves scanning along one of the spatial dimensions to estimate the depth of an scene using structured light, or the use of two cameras as required by stereo-imaging techniques. Recently, Time-of-Flight (ToF) snapshot imaging has gained considerable attention, due to its accuracy and speed. To date, however, these imaging modalities (CSI and ToF) have been realized and implemented by separate independent imaging sensors. This paper presents the development of a single aperture compressive spectral + depth imaging camera that employs a commodity 3-D range ToF sensor as the sensing device of a coded-aperture-based compressive spectral imager. The proposed system uses a single aperture/single sensor; thus, representing a significant improvement over existing RGB+D cameras that integrate two separate image sensors, one for RGB and another for depth. In addition, the observable wavelength range of the CSI device is expanded from the visible to the near-infrared, resolving up to as many as 16 independent channels. The proposed system allows the addition of side-information by placing a grayscale or RGB camera in the same path of the single-aperture system, such that the quality of the spectral estimation is improved, while maintaining high-frame rates. We demonstrate the proposed ideas through real experimentation conducted on an assembled CSI+ToF testbed camera system.**

***Index Terms*—Compressed sensing, depth estimation, RGB+D imagery, spectral imaging, time-of-flight imaging.**

I. Introduction

IMAGING spectroscopy, commonly referred to as multispectral or hyperspectral imaging is the process of sensing large amounts of spatial information across a multitude of wavelengths [1]. Compressive spectral imaging (CSI) systems comprise the new generation of spectral imagers, since they record all wavelengths of a scene in a single snapshot, simultaneously [1]. In CSI systems, a ray of light emanating from the scene is passed through a series of optical elements that spread the single ray of light over a lateral sequence of sensor pixels, in a wavelength dependent manner, just like a spectrophotometer would spread light across a linear array. Allowing multiple rays of light onto the sensor means that the camera can resolve a two dimensional image; however, because the spreading of light rays across lateral sequences of pixels means that a single pixel will collect light from multiple sources, compressive sensing techniques are used to decouple the spectral profiles of neighboring pixels [2], [3].

On the other hand, 3-D range imaging, refers to the process of recording a 2-D image of a scene where each pixel is a measure of the distance from the camera sensor to the target surface. There exist two main modalities to estimate 3D imagery: passive illumination such as stereo [4] and light field imaging [5], [6], and active illumination such as structured light imaging (SLI) [7], [8] and time-of-flight (ToF) imaging [9], [10]. One of the big disadvantages of passive systems is that they cannot derive depth without depth cues created by texture; this is known as the depth stereo problem. Passive methods and SLI, rely on triangulation to interpolate depth values in the spaces between edges. The problem with triangulation is the difficulty to work over long ranges since the size of the triangle has to grow proportionally. They also have to deal with problems of occlusion where target surfaces are visible to the camera but not to the projector. ToF range cameras, in contrast, indirectly measure the round-trip propagation delay of the light pulse by strategically triggering each pixel's exposure time such that it captures a portion of the projected pulse. ToF sensors do not suffer of occlusion problems, since the projector and camera have the same line of sight. Moreover, ToF sensors can reach longer ranges by just changing the frequency and amplitude of the modulated light, they are lightweight, power efficient, and can collect full frame images with a single laser pulse. Sensors from Mesa Imaging, Microsoft, SoftKinetic, PMD Tech, among others, rely on custom imaging chips that combine traditional CCD and CMOS light detectors with custom application specific integrated circuits (ASICs). While effective, these techniques tend to produce camera sensors with low pixel counts.

In some cases, the information in both imaging modalities can be combined as in RGB+Depth imaging to infer qualitative features in a scene [11], [12]. To date, however, these

Manuscript received December 1, 2016; revised May 5, 2017 and June 16, 2017; accepted July 20, 2017. Date of publication August 9, 2017; date of current version October 23, 2017. This work was supported under Grant Intel/NSF VEC 1538950. The work of H. Rueda was supported by a Colciencias-Fulbright scholarship. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Yebin Liu. *(Corresponding author: Hoover Rueda.)*

H. Rueda, C. Fu, and G. R. Arce are with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 USA (e-mail: rueda@udel.edu; fuchen@udel.edu; arce@udel.edu).

D. L. Lau is with the Department of Electrical and Computer Engineering, University of Kentucky, Lexington, KY 40506 USA (e-mail: dllau@uky.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2737784

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

imaging modalities have been realized and implemented by separate independent imaging sensors.

A. Related Work

The importance of spectral+3D imaging has been reported recently. Conventional joint spectral and depth imaging usually requires scanning the 4-D scene either spatially or spectrally. Latorre-Carmona, *et. al.* [13] proposed to combine a lenslet-based integral imaging with a liquid crystal tunable filter that scans each spectral band for spectral integral imaging. Spectral scanning can be avoided through combining a lenslet array with a pixel-wise color filter [14], [15]; however, the spectral resolution is limited by the number of colors in the pixel-wise filter. To overcome scanning and the spectral resolution limitation, compressive spectral integral imagers have been developed employing a liquid crystal or a dispersive element for spectral modulation [16], [17]. These compressive imagers only apply compressive sensing to acquire spectral elemental images. Thus, a second-step integral imaging reconstruction must be followed. Recently, snapshot spectral integral imagers based on lenslet arrays were introduced, where spectro-volumetric images are directly reconstructed from compressed projections [18], [19]. A different approach obtains the spectral and depth information of a scene by adding a grayscale camera on the side of a multi-spectral imager [20], such that a passive stereo system is attained.

In spite of the above 4-D imagers employing passive range finding techniques, active ranging such as laser scanning methods were also applied in spectral imaging. M. Kim, *et al.* [21] combines a coded aperture snapshot spectral imager with a conventional laser scanner in their spectral imager design. This imaging system, however, involves a scanning illumination source and two independent detectors. As a replacement of the laser scanning based ranging, ToF cameras acquire the depth information simultaneously on the sensor array. Compressive sensing has been applied with ToF techniques for high resolution depth map reconstructions [22], [23].

B. Contribution

To our knowledge, spectral imaging based on ToF sensors has been limited to combining a ToF camera with a separate spectral imager or an RGB camera [24]. This paper introduces a new spectral+TOF imaging camera that measures and reconstructs multispectral and 3D imagery at once. It aims to create superhuman sensing that will enable the perception of our world in new and exciting ways. We intend to advance in the state of the art in compressive sensing systems to extract depth while accurately capturing spectral material properties. The applications of such a sensor are self-evident in fields such as computer/robotic vision because they would allow an artificial intelligence to make informed decisions about not only the location of objects within a scene but also their material properties. These goals are made possible after the realization that ToF sensors can exploit both ambient light and modulated light, in order to recover pixelated depth and multispectral information. In particular, we build on this opportunity and integrate the capabilities of a CMOS ToF sensor into a CSI snapshot camera. So unlike a Kinect 2.0 or Prime Sense RGB+D camera, the proposed system uses a single aperture image sensor to capture multispectral+depth (MS+D) images. Moreover, by changing the dispersive element, used in the proposed system, by a diffraction grating, we can move towards hyperspectral+depth (HS+D) imagery, as it will be described shortly.

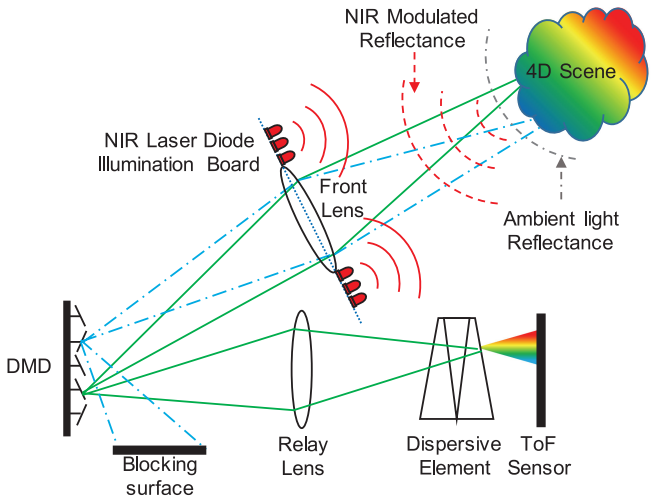


Fig. 1. Optical design sketch of the single aperture snapshot ToF+Spectral camera. The LED laser diode illumination board is attached to the front lens of the camera which images the 4D scene onto the DMD. Depending on the mode, the DMD will emulate an all-on or a random coded aperture, for the ToF and ambient light capture, respectively. The relay lens transmits the DMD coded and modulated light through a dispersive element, so that monochrome coded, modulated and sheared light is integrated on the ToF sensor.

II. Snapshot Single Aperture Compressive ToF+Multispectral Camera

The ability to switch measurement modes to capture both ambient light and ToF depth images in the same sensor array enables a dual-shot single aperture imaging system, where spectral and 3D range depth images are reconstructed from the same ToF sensor. Fig. 1 depicts a sketch of the proposed single aperture system. It uses a set of lenses for image formation, an LED laser diode illumination board for ToF depth excitation, a digital micromirror device (DMD) to emulate the coded apertures, a dispersive element to spread the ambient illumination, and a ToF sensor to integrate the incoming light field, which can operate in ToF or ambient light mode. The depth image is obtained through processing four discrete pulsed frames measured in the ToF mode, as will be explained next, by turning on all the mirrors of the DMD, while the spectral images will be recovered from compressed projections at the ambient grayscale mode.

A. Time-of-Flight Imaging (ToF)

Time-of-flight imaging aims at acquiring the depth image or depth map of the underlying scene using an array of modulated infrared laser diodes as an illumination source. The time delay of the returned light for each spatial location is measured on a two-dimensional (2D) ToF sensor array. According to the laser modulation, there are typically two ways of measuring

this time delay [9], [10]. One of them directly measures the round-trip traveling time of the shinned laser light pulses by detection of the first arriving laser photons. Thus the depth value is calculated by $d = c\frac{t}{2}$, where c is the speed of light constant. In this method, the depth measurement accuracy depends on the time measurement resolution. Another ToF imaging method employs a continuous sinusoidal or square-wave laser modulation. Instead of directly measuring the time delay, this method relates the depth information to the phase difference between the emitted and reflected light. In this method, the depth accuracy is determined by the phase measuring capability at a given modulation frequency. Note that the modulation frequency used entails a trade off between depth measurement capability and accuracy; lower frequencies can measure depth at longer distances but with reduced accuracy.

This work uses an Espros EPC660 ToF imaging sensor, where continuous sinusoidal laser modulation is used by default [25]. In phase-based ToF sensors like the Microsoft Canesta [26] or Espros EPC660, distance is measured indirectly as the round-trip phase delay between a modulated light source and synchronized array of gated pixels. To derive a mathematical model for these ToF sensors, we begin by defining $g_{m,n}$ as a single pixel of a traditional $M \times N$ sensor array with discrete row and column coordinate, (m, n), such that that amount of light accumulated over an exposure, τ, is defined according to:

$$g_{m,n} = \sum_{l=0}^{L-1} \frac{1}{\tau} \int_{t=0}^{\tau} A f_{m,n,l} dt = \sum_{l=0}^{L-1} A f_{m,n,l} \tag{1}$$

where A represents the amount of ambient light present inside the field of view, and $f_{m,n,l}$ is the reflectance of that light from the target surface, as a function of wavelength l. Adding a modulated light source $B\cos(2\pi ft)$, the captured pixel becomes:

$$g_{m,n} = \sum_{l=0}^{L-1} \frac{1}{\tau} \int_{t=0}^{\tau} \left(A f_{m,n,l} + B f_{m,n,l_0} \cos(2\pi ft + \theta)\right) dt, \tag{2}$$

where θ is the phase delay introduced by the round-trip travel time of the modulated light from source to target and back, and f_{m,n,l_0} is the surface reflectance at the wavelength of the modulated light source (l_0). In our single aperture system shown in Fig. 1, the laser illumination board will illuminate the 4D scene with the modulated light, and the reflected signal from the scene will propagate through the optical path of the imaging system as shown in Fig. 2(a). Note that all the mirrors of the DMD will be set to on, such that all reflected light from the scene propagates to the sensor.

To represent the gating of pixels, we define the gating function, $r(t)$, as a square wave in the range, $[0, 1]$, with a frequency, f, matched to the modulated light source and with a duty cycle D such that:

$$g_{m,n} = \sum_{l=0}^{L-1} \frac{1}{\tau} \int_{t=0}^{\tau} \left(A f_{m,n,l} + B f_{m,n,l_0} \cos(2\pi ft + \theta)\right) r(t) dt. \tag{3}$$

Denoting the total exposure time as K times the period of one wavelength w (in seconds) of the square wave, then $\tau = K * w$.

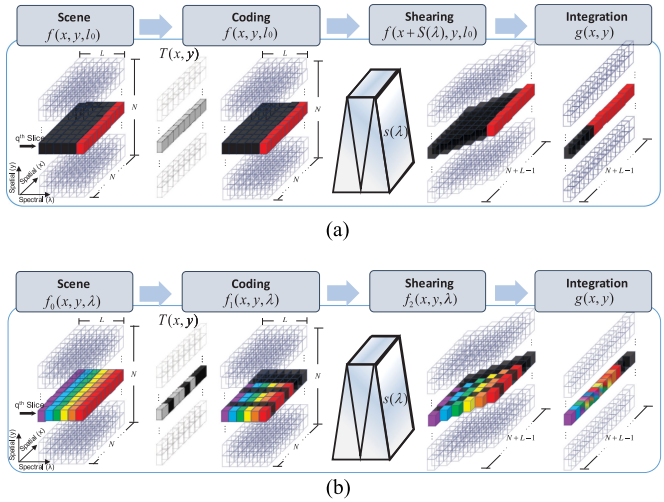


Fig. 2. Light propagation of the q^{th} slice through the proposed system. (a) In ToF mode, the reflected NIR modulated light is captured after ambient light is subtracted. The single red spectral band represents the reflected NIR light of the LED illumination. The DMD emulates an all-transmissive pattern, such that, no coding is performed. Therefore, propagation through the system causes no multiplexing but shifting in the captured image. (b) In ambient light mode (CSI mode), the incoming light is encoded by the coded aperture pattern $T(x, y)$, then the dispersive element decomposes the light in the corresponding wavelengths, and multiplexing takes place at the sensor. (a) ToF operation mode. (b) CSI operation mode.

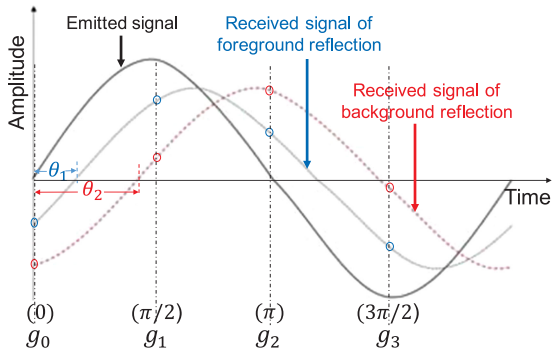


Fig. 3. ToF sensor raw data g_0 to g_3 are captured in every single $(m, n)^{th}$ pixel. The corresponding phase delays θ_0 to θ_3 are then calculated to obtain the depth map.

Given that, $f = 1/w$, then $f = K\tau^{-1}$, such that there are an integer, K, periods of $\cos(2\pi ft)$ and $r(t)$ inside of τ. Therefore, by integrating over one period instead of all of τ, we simplify (3) to:

$$g_{m,n} = \sum_{l=0}^{L-1} \frac{K}{\tau} \int_{t=0}^{\frac{D\tau}{K}} \left(A f_{m,n,l} + B f_{m,n,l_0} \cos(2\pi ft + \theta)\right) dt. \tag{4}$$

Assuming a small D, $g_{m,n}$ can be approximated as:

$$g_{m,n} \approx \sum_{l=0}^{L-1} DA f_{m,n,l} + DB f_{m,n,l_0} \cos(\theta). \tag{5}$$

By itself, there is no way to extract θ separate from B, but if, as demonstrated in Fig. 3, we collect a sequence of 4 images indexed by ℓ with a phase shift between the modulated light

source and camera images such that

$$(g_\ell)_{m,n} \approx \sum_{l=0}^{L-1} DAf_{m,n,l} + DBf_{m,n,l_0} \cos\left(\frac{\pi\ell}{2} + \theta\right), \quad (6)$$

then we can define the difference of complements, $g_{0,2}$ and $g_{3,1}$, according to:

$$\begin{aligned}(g_{0,2})_{m,n} &= (g_0)_{m,n} - (g_2)_{m,n}, \\ &= 2DBf_{m,n,l_0} \cos\theta \end{aligned} \quad (7)$$

and

$$\begin{aligned}(g_{1,3})_{m,n} &= (g_1)_{m,n} - (g_3)_{m,n}, \\ &= 2DBf_{m,n,l_0} \sin\theta \end{aligned} \quad (8)$$

and, finally, extract θ by treating $(g_{0,2})_{m,n}$ and $(g_{1,3})_{m,n}$ as quadrature components such that:

$$\begin{aligned}\theta &= \tan^{-1}\left(\frac{(g_{1,3})_{m,n}}{(g_{0,2})_{m,n}}\right), \\ &= \tan^{-1}\left(\frac{2DBf_{m,n,l_0}\sin\theta}{2DBf_{m,n,l_0}\cos\theta}\right), \\ &= \tan^{-1}\left(\frac{\sin\theta}{\cos\theta}\right). \end{aligned} \quad (9)$$

The depth value, at every $(m,n)^{th}$ sensor pixel, $d_{m,n}$, is then calculated as

$$d_{m,n} = \frac{c}{2}\frac{\theta}{2\pi f}, \quad (10)$$

where, for the case of the Espros EPC660 sensor, the laser modulation frequency, f, can be selected from 0.625 MHz to 24 MHz, resulting in an unambiguous distance ranging of up to $d_{m,n} = 6.25$ meters for $f = 24$ MHz and up to $d_{m,n} = 240$ meters for $f = 0.625$ MHz.

Separate from range, we can extract the surface reflectance from the modulated light, f_{m,n,l_0}, by looking at the magnitude of the quadrature components such that:

$$f_{m,n,l_0} = \frac{\sqrt{(g_{1,3})^2_{m,n} + (g_{0,2})^2_{m,n}}}{2DB}, \quad (11)$$

given that $2DB$ is known a priori. Alternatively, to extract the surface reflectance under ambient light illumination, we can either use $(g_0)_{m,n}$ and $(g_2)_{m,n}$ which are 180° out-of-phase, as are $(g_1)_{m,n}$ and $(g_3)_{m,n}$, by evaluating their sum,

$$(g_0)_{m,n} + (g_2)_{m,n} = 2\sum_{l=0}^{L-1} DAf_{m,n,l} \quad (12)$$

such that:

$$\sum_{l=0}^{L-1} Af_{m,n,l} = \frac{(g_0)_{m,n} + (g_2)_{m,n}}{2D}. \quad (13)$$

This last equation demonstrates that phase-based ToF sensors like the Espros EPC660 have an inherent ability to separately measure both ambient (DC) and modulated (AC) light images, and given this inherent ability, the integration of ToF sensors into a CSI system becomes almost compulsory since the unused DC image can now be used for multispectral imaging while the modulated image reserved for ToF.

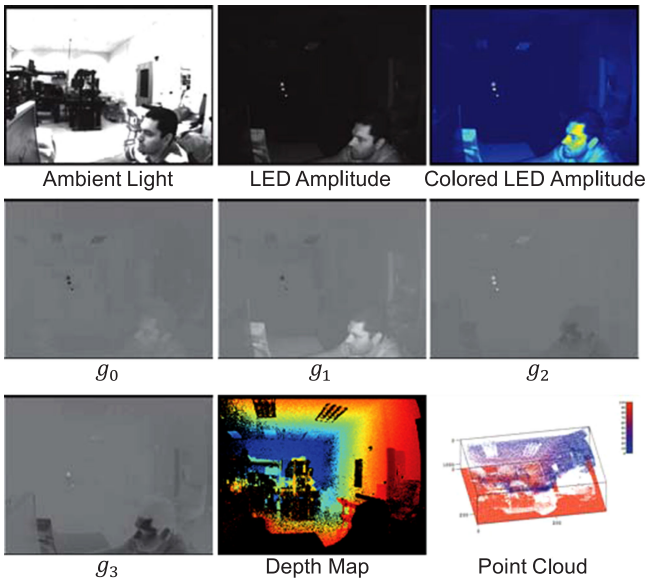


Fig. 4. Ambient image is recorded in grayscale mode. Four successive frames g_0, g_1, g_2 and g_3 are captured for the depth map and point cloud estimation.

To illustrate the entire process, Fig. 4 shows the 9 different images provided by the ToF sensor. The original ambient light image of (2) is included as the (top-left) grayscale image. The amplitude of the quadrature components in (11) is illustrated as a (top-center) grayscale and (top-right) pseudo-color image. The sequence of phase shifted images of (6) are illustrated as signed grayscale images as the sequence from (middle-left) to (bottom-left). Finally, the depth estimation is illustrated as a (bottom-center) pseudo-color and (bottom-right) point cloud.

B. Compressive Spectral Imaging

Compressive spectral imaging (CSI) encompass techniques which sense the spatio-spectral information of a scene by using 2D coded projections, and as few as a single coded projection. The underlying spectral 3D data cube is then recovered using compressive sensing (CS) reconstruction algorithms which assume that the hyperspectral images are sparse in some representation basis. The advantage of CSI is that the required number of measurements needed for reconstruction is far less than that required by traditional scanning methods [27]–[29]. In the CSI literature different strategies to attain the coded projections have been proposed. The review articles in [1], [30]–[32] provide a number of architectures that attain coded projections. One of the main families in the state-of-the-art CSI architectures are those based on coded apertures. These use a set of lenses, a coded aperture as coding element, a dispersive element and a 2D focal plane array (FPA) sensor, to capture compressive multiplexed random projections of the scene.

As depicted in Fig. 2(b) the front-end lens in these architectures effectively creates a light field in front of the coded aperture plane, such that $f_0(x, y, \lambda)$ represents the light ray

incident upon the coded aperture at 2-D coordinate (x, y) and having wavelength λ. As light approaches the sensor, the coded aperture blocks or allows the rays of light from reaching the sensor. This aperture effect can be accounted in our model by means of the binary modulation function $T(x, y) = \sum_{m,n} T_{m,n} \text{rect}\left(\frac{x}{\Delta_c} - n, \frac{y}{\Delta_c} - m\right)$, where $T_{m,n} \in [0, 1]$. In order to separate the individual wavelengths of the incoming signal, the system places a dispersive element in front of the image sensor such that $f_1(x, y, \lambda)$ is split into discrete bands spread across L consecutive pixels along the lateral or n dimension in a wavelength dependent manner, indicated by the dispersion function $S(\lambda)$. Accumulating all of the light incident upon a discrete pixel of the sensor with row and column coordinate $[m, n]$ and defining $\chi_{m,n,l}$ as the spectral sensitivity of the sensor for that same discrete pixel, the monochrome pixel, $g_{m,n}$, of the resulting digital image is given by:

$$g_{m,n} = \sum_{l} \chi_{m,n,l} T_{m,n-l} f_{m,n-l,l} + \omega_{m,n} \tag{14}$$

where $l = 0, \ldots, L-1$ indexes the pixels affected by the spreading of $f_1(x, y, \lambda)$ as a result of $S(\lambda)$. One would expect $\chi_{m,n,l}$ to be either constant for all pixels and absorb light in the visible light range from 350 to 750 nm or selectively absorb red, green, or blue light in a traditional Bayer pattern.

In order to establish a matrix representation consistent with CS, we define $\mathbf{f}$ as a column vector containing $f_{m,n,l}$ for all spectral components l and all pixels m and n ordered first by row, then column, and finally by wavelength. Assuming we have a coded aperture with $M \times N$ pixels ($\mathbf{T} \in \mathcal{R}^{M \times N}$), $\mathbf{f}$ will be MNL in length ($\mathbf{f} \in \mathcal{R}^{MNL}$). Then, we define $\mathbf{g}^k$ as a column vector holding all of the recorded pixel values, $g_{m,n}$, from the monochrome image sensor for the k^{th} image in our sequence of $K \leq L$ snapshots. Multiple snapshots allow to obtain different information from the same scene as different coded patterns can be used. Noting the lateral dispersion of light, the length of $\mathbf{g}^k$ is $M(N + L - 1)$, such that all light is accounted for on the sensor. So while the coded aperture has an $M \times N$ pixel array, the sensor has an extra set of $L - 1$ pixel columns. Assuming that $\mathbf{f}$ stays constant over the K snapshots, we can relate $\mathbf{g}^k$ to $\mathbf{f}$ according to:

$$\mathbf{g}^k = \mathbf{P}\mathbf{X}\mathbf{T}^k\mathbf{f} = \mathbf{H}^k\mathbf{f}, \tag{15}$$

where $\mathbf{H}^k$ represents the combined effects of the dispersion effect $\left(\mathbf{P} \in \mathbb{R}^{M(N+L-1) \times MNL}\right)$, the sensor pixel-wise spectral sensitivity in matrix form $\left(\mathbf{X} \in \mathbb{R}^{MNL \times MNL}\right)$, and the block-diagonal representation of the k^{th} coded aperture $\left(\mathbf{T}^k \in \mathbb{R}^{MNL \times MNL}\right)$.

Alternatively, the sensing process can be expressed as $\mathbf{g} = \mathbf{H}\mathbf{f} = \mathbf{H}\boldsymbol{\Psi}\boldsymbol{\theta} = \mathbf{A}\boldsymbol{\theta}$, such that $\mathbf{g} = \left[(\mathbf{g}^0)^T, \ldots, (\mathbf{g}^{K-1})^T\right]^T$, $\mathbf{H} = \left[(\mathbf{H}^0)^T, \ldots, (\mathbf{H}^{K-1})^T\right]^T$, $\mathbf{A} = \mathbf{H}\boldsymbol{\Psi}$ is the compressive sensing matrix, and $\boldsymbol{\theta} = \boldsymbol{\Psi}^T\mathbf{f}$ is a vector coordinate of $\mathbf{f}$ in an orthogonal basis $\boldsymbol{\Psi}$. Conventional basis functions include the Wavelet ($\boldsymbol{\Psi}_{W\,2D}$) and discrete cosine transforms, ($\boldsymbol{\Psi}_{DCT}$) as well as pre-trained dictionaries [33], and bases which simultaneously exploit the sparsity properties along a multidimensional space by doing the Kronecker product of bases used for each of its dimensions [34]–[36]. The basis transformation assumes $\mathbf{f}$ to be sparse on $\boldsymbol{\Psi}$, so that, a small subset of the basis vectors can accurately reconstruct $\mathbf{f}$ with small distortions. An estimation of the spectral input data cube can be attained by solving the regularization problem,

$$\hat{\mathbf{f}} = \boldsymbol{\Psi}\left(\underset{\boldsymbol{\theta}'}{\text{argmin}} \|\mathbf{g} - \mathbf{A}\boldsymbol{\theta}'\|_2^2 + \tau\|\boldsymbol{\theta}'\|_1\right), \tag{16}$$

where $\boldsymbol{\theta}'$ is the estimation of the sparse coefficients, $\|\cdot\|_2^2$ represents the ℓ_2 norm, $\|\cdot\|_1$ the ℓ_1 norm, and τ is a regularization parameter which penalizes the fact of $\boldsymbol{\theta}'$ being sparse, while reducing the error between the estimated $\mathbf{A}\boldsymbol{\theta}'$, and the captured set of compressive measurements $\mathbf{g}$. A large variety of solvers for problem (16) has been proposed, including, the two-step iterative shrinkage/thresholding (TwIST) using the total-variation (TV) as the regularization function [37], the gradient projection for sparse reconstruction (GPSR) [38], Gaussian mixture models (GMM) [39], and denosing algorithms based on approximate message passing (AMP) [40].

C. Joint Compressive Spectral+ToF Imaging

The previous subsections described the reconstruction of spectral and 3D ranging imagery, performed separately and independently. However, the correlations between the spectral bands and the ToF quadrature components g_0 to g_3, encourages the joint reconstruction of spectral+depth imagery. In addition, side information can be used to improve the reconstruction quality as demonstrated in [20], [41], [42]. The collection of side information usually involves mounting an additional grayscale [20], [41] or RGB camera [42], either side-by-side or along the same path of the multi-spectral system by means of a beam-splitter. The reconstruction algorithm then exploits the correlation between the multi-spectral measurement and the grayscale, or RGB side image. In the proposed single-aperture system, and as a first approach, side information can be extracted from the ToF quadrature components with no additional camera needed. In particular, the surface reflectance of the target scene at the LED modulation wavelength (l_0) can be calculated according to (11). The extracted monochrome image can then be viewed as the ground truth of the scene at l_0-nanometers. Mathematically, we assume that the spectrum of the collected surface reflectance falls into the range of the last reconstructed spectral band. Denote $\mathbf{f}_r$ as the vector form of the collected image band containing the side information. Then the joint sensing process for spectral imaging can be reformulated as,

$$\begin{bmatrix} \mathbf{g} \\ \mathbf{f}_r \end{bmatrix} = \begin{bmatrix} \mathbf{H} \\ \mathbf{H}_b \end{bmatrix} \mathbf{f}, \tag{17}$$

where $\mathbf{H}_b = [\mathbf{0}, \mathbf{I}]$, such that $\mathbf{0}$ is an $MN \times MN(L-1)$ zero matrix and $\mathbf{I}$ is an $MN \times MN$ identity matrix.

In addition, since compressive sensing reconstruction algorithms tend to converge to smoothed estimations, the edge information from the ToF amplitude image can also be incorporated as side information across the spectrum, since each spectral band preserves most of the borders in the reflectance image. To this end, an edge detector is first applied to the modulated surface

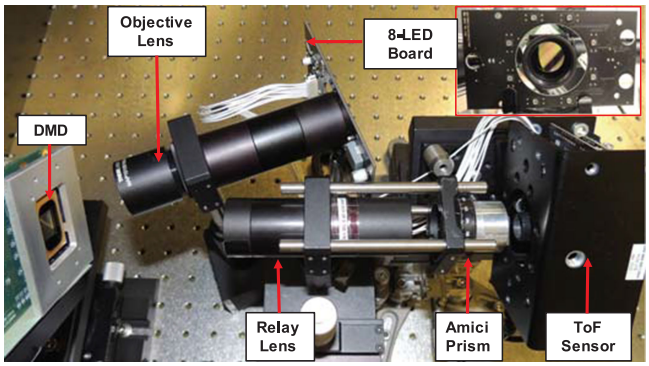


Fig. 5. Compressive single-aperture ToF+CSI testbed. An objective lens with an attached LED illumination board images the light field of the target scene onto the DMD. After the DMD, a VIS-NIR relay lens transports the coded light through the dispersive element, which decomposes the light field in the corresponding wavelengths being integrated in the ToF sensor device. Side information can be obtained by incorporating a beam-splitter and a secondary sensor as proposed in [41].

reflectance image to extract a binary edge image $\mathbf{E}$. Then, by noting that the grayscale values of the edge pixels vary with the spectrum wavelengths, the edge image needs to be scaled by means of a calibration data cube captured at each specific wavelength $\mathbf{C}_\ell$. This calibration data cube needs to be captured only once, and off-line, so that it does not affect the frame-rate of the proposed system. Mathematically, the side-information from the edge image can be incorporated in the sensing process, by rewriting (17) as, $[\mathbf{g}^T, \mathbf{f}_e^T]^T = [\mathbf{H}^T, \mathbf{H}_e^T]^T \mathbf{f}$, where $\mathbf{H}_e = [\mathbf{H}_E, \mathbf{I}]$, with $\mathbf{H}_E$ being the concatenation of $L-1$, $MN \times MN$ diagonal matrices, with elements being the vectorized point-wise multiplication between the edge image $\mathbf{E}$ and the corresponding calibrated image $\mathbf{C}_\ell$; $\mathbf{f}_e$ is the estimated grayscale image with the scaled border information.

Side information via a grayscale or an RGB camera can be incorporated at the input of the single-aperture proposed system by use of a beam-splitter. The beam-splitter allows the same path to be shared by the two cameras, so that, registration between the 2 images can be easily achievable. Note that by including this additional camera to the proposed single-aperture ToF architecture, higher flexibility is attained since CSI+ToF+side information can now be jointly captured. Mathematically, this approach to side-information sensing, can be written as $[\mathbf{g}^T, \mathbf{g}_s^T]^T = [\mathbf{H}^T, \mathbf{H}_s^T]^T \mathbf{f}$, where $\mathbf{g}_s \in \{\mathbf{g}_g, \mathbf{g}_c\}$, such that $\mathbf{g}_g \in R^{MN \times 1}$ is the captured grayscale side image in vector form, and $\mathbf{g}_c \in R^{3MN \times 1}$ is the RGB side image in vector form. In addition, $\mathbf{H}_s \in \{\mathbf{H}_g, \mathbf{H}_c\}$, where $\mathbf{H}_g$ is the concatenation of L, $MN \times MN$ all-one diagonal matrices representing the grayscale side sensing matrix, and $\mathbf{H}_c$ represents the RGB side-information sensing matrix, which concatenates L, $3MN \times MN$ diagonal matrices. The entries of $\mathbf{H}_c$ correspond to the discrete response of the filters from the Bayer pattern of the RGB camera.

Side information can also be used to improve the depth map estimation. Typical denoising techniques can be used to smooth the depth imagery, but filtering tends to generate smooth images with blurred target borders. Note however that object borders can be easily obtained either from the quadrature components, the grayscale or RGB side information. With this prior border information, a center weighted median filter (CWMF) [43] can then be applied to the initial depth map. The advantage of using CWMF is that a median filter is applied only on surfaces inside the borders, while the values on the borders are unmodified. This, in turn, guarantees a smoother depth estimation without blurring edges.

Other possible approaches to process the ToF surface reflectance image as side information in spectral imaging include extracting a sparse spatial basis through dictionary learning, as described in [20]. Joint reconstruction of compressive spectral and depth imagery is a promising topic which requires more future research for significant improvements on the reconstruction quality in both spectral and depth dimensions.

III. Compressive ToF+Multispectral Testbed Imaging System

The proof-of-concept optical prototype presented in Fig. 5 was assembled in our lab. In this system, the imaging arm is composed of an 8-LED laser diode board attached to a VIS-NIR matched achromatic doublet objective lens, and a digital micromirror device; a beam-splitter can be used to enable side information from a secondary sensor as proposed in [41]. The integration arm is composed of a VIS-NIR doublet as the relay lens, a custom VIS-NIR double Amici prism as the dispersive element, and an Espros EPC660 ToF sensor. A target scene is illuminated with a broadband illumination lamp source, and with the modulated LED light with frequencies ranging between $f = 0.625$ MHz and $f = 24$ MHz. The reflected light from the scene is captured by the objective lens and focused onto the mirrors of the DMD image plane, which plays the role of the coded aperture, and onto the secondary sensor, if the beam-splitter is present. When properly aligned, the mirrors of the DMD reflect light into the integration arm, which relays light through a second lens and then through the prism. The dispersed field focuses in the ToF image plane, which imposes its spectral filtering and integration based on its wavelength sensitivity that spans the range between 450 and 1000 nm.

The DMD pixel size is 13.68 um, and the size of the ToF sensor pixels is 20 um. A 3×3 macro-pixel windowing is done in the DMD, and a 2×2 macro-pixel windowing is realized on the ToF sensor, such that an approximately one-to-one correspondence between the two elements (DMD macro-pitch $=$ 41.04 um, and ToF macro-pitch $=$ 40 um) is attained. In order to have dyadic dimensions to perform Wavelet-based representation recovery, and given the DMD and ToF sensor resolution, the coded aperture patterns exhibit a resolution of 128×128 pixels (384×384 DMD pixels, and 256×256 ToF sensor pixels). More on resolution improvements will be discussed in Section IV.

The prism used in the testbed is a custom designed non-linear double Amici prism, which disperses the wavelength spectrum between 486–766 nm onto 16 sensor pixels. If a photomask or a DMD with the same pixel size as that of the sensor is used instead, we would be able to recover 16 spectral band datacubes. However, we are able to recover up to 8 spectral

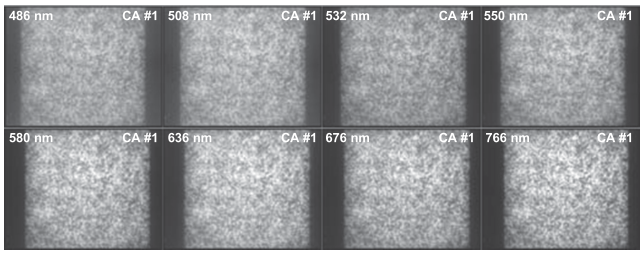
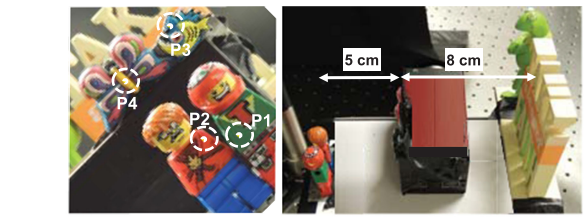
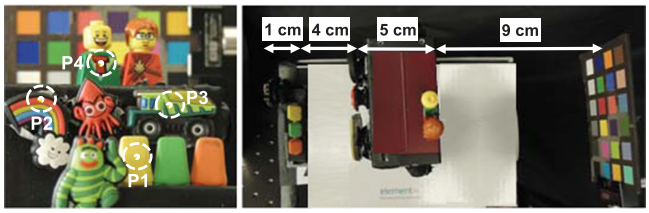


Fig. 6. Coded aperture calibration patterns as seen by the ToF sensor when illuminated with monochrome light at the given wavelengths. The coded aperture is the result of a binary random variable with 50% transmittance.



(a)



(b)

Fig. 7. Target scenes used in the experiments. The first scene is placed around 40 centimeters away from the objective lens, and the second scene is placed around 50 centimeters away. Spectral signatures of the four highlighted points (P1-P4) will be compared in Figs. 12 and 13. (a) Target scene 1. (b) Target scene 2.

channels, due to the 2×2 macro-pixel windowing. The cut-off wavelengths of these 8 channels are 486, 508, 532, 550, 580, 636, 676, and 766 nanometers. The bandwidth of the spectral channels is non-uniform due to the non-linearity of the prism. The attained reconstruction with the testbed will therefore exhibit $N \times N = 256 \times 256$ pixels of spatial resolution and $L = 8$ spectral channels of spectral resolution. Note that, the spectral resolution can be increased significantly by the use of a diffraction grating.

A. Calibration of the System

To test the multispectral capabilities of our system, we use the grayscale mode of the ToF sensor in order to capture the raw ambient light data. Each coded aperture used, should be calibrated to determine the experimental impulse response of the system at each wavelength. The calibration is done using a high-reflective white plate as the target scene, illuminating it with monochromatic light at the cut-off wavelengths of the dispersive element. The monochromatic light is obtained from a Tunable Xenon Arc Light source, which provides nanometer resolution. A calibration data cube is thus captured for each coded aperture. Since the testbed setup can reconstruct up to 8 spectral channels, we can capture up to 8 snapshots to estimate

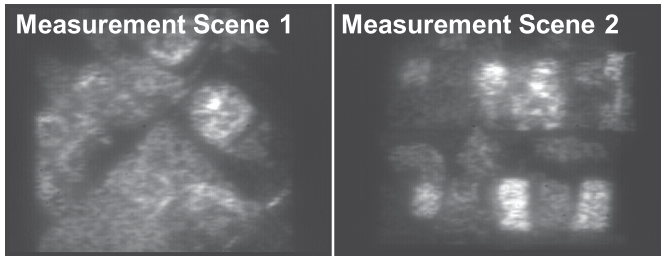
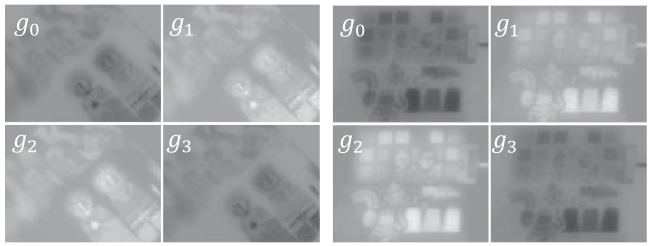


Fig. 8. Compressive measurements as seen by the ToF sensor when the target scenes are illuminated with broadband fluorescent white light and the DMD exhibit a random coded pattern.



(a) (b)

Fig. 9. Sequence of four images g_0, g_1, g_2, g_3 captured with the testbed, in ToF mode using $f = 12$ MHz, for (a) Scene 1, and (b) Scene 2.

the underlying input scene, while still satisfying compression. Therefore, we analyzed the use of $K = 1$, $K = 2$, and $K = 4$ snapshots for the spectral reconstructions, each one using a different realization of a random coded aperture with 50% of transmittance (50% of the elements are 1, and the other are 0). The calibration data cube for 1 of the 4 coded apertures used in the experiments is shown in Fig. 6; each image represents the PSF attained on the ToF sensor at the corresponding monochromatic illumination wavelength. These calibrated coded apertures are then used in the reconstruction algorithm as the $\mathbf{H}$ matrix.

B. Capturing Compressive Spectral and ToF Measurements

Compressive measurements of a real target use a broadband white illumination source instead of the monochromatic light used in the calibration. To test the efficiency of the images attained with the testbed, we evaluate 2 target scenes, depicted in Fig. 7, placed at around 40 and 50 centimeters away from the front-end lens, respectively. The first target scene exhibits 3 depth planes within a range of 13 centimeters, whereas the second scene exhibits 4 depth planes within a range of 20 centimeters.

Fig. 8 shows an example of the compressive measurements attained with the testbed setup, under the ambient light reading mode, for the two target scenes. Note that the different snapshots integrate approximately the same amount of intensity, restricted by the transmittance of the coded aperture, which was fixed to be 50%. Also note that, the higher the number of snapshots, the higher quality in the reconstruction is expected.

Similarly, in the ToF reading mode, we turn on all the mirrors of the DMD, the LED laser diode illumination board shines the sinusoidal function as explained in Section II-A, the broadband illumination source remain on, and the system then captures

the four consecutive frames g_0, g_1, g_2 and g_3 in a single period of exposure time. Fig. 9 shows the 4 frames captured using a modulation frequency of $f = 12$ MHz, for the 2 target scenes. In this figure, it can be noticed the reciprocity between g_0 and g_2, and g_1 and g_3.

C. Reconstruction Results for CSI

To reconstruct the spectral data cubes from the coded projections, the compressive-sensing-based gradient projection for sparse reconstruction (GPSR) algorithm [38] was used to solve the optimization problem in (16), with and without the use of side information. In our experiments, we set $\mathbf{\Psi} = \mathbf{\Psi}_{W\,2D} \otimes \mathbf{\Psi}_{DCT}$ to be the Kronecker product of the 2D Symlet 8 Wavelet transform and the 1D DCT. The grayscale or RGB side information images are emulated by following the same procedure used in the calibration, replacing the white plate with the corresponding scene, and the DMD exhibiting a full-transmissive pattern. We then register and map the attained data cube to the spectral responses $\chi_{m,n,l}$ of the Stingray F033B and F033C cameras, to generate the 256×256 grayscale, and $256 \times 256 \times 3$ color side images.

Fig. 10 presents the reconstructed data cubes, for the two target scenes in Fig. 7, using $K = 1, 2$ and 4 snapshots, mapped to an RGB profile, to test the spatial fidelity. It can be noticed that the reconstruction quality improves with additional snapshots from the single-aperture system, or by incorporating side-information. Note that although the ToF side information impacts the final results, the grayscale and RGB side information have a higher impact with as low as a single snapshot.

The performance gain achieved with the use of side information is illustrated in Fig. 11, where 4 out of the 8 reconstructed spectral channels are shown, when $K = 1$ is used. The corresponding band/color discrimination can be contrasted with the RGB reconstructions presented in Fig. 10. For instance, for scene 1, the green chest of the target on the right is well characterized at the 532 nm wavelength, whereas its red helmet clearly appears at the 676 nm wavelength. For scene 2, the yellow eraser at the bottom is well reconstructed at 580 nm, the red octopus is discriminated by the 676 nm wavelength, and the green chest of the Lego toy is again well characterized at the 532 nm wavelength. Here, it can be confirmed the boost in quality entailed by the use of side information.

To evaluate the spectral reconstruction results, Figs. 12 and 13 present the spectral footprint of 4 different points of the 2 target scenes, respectively, denoted as P1, P2, P3 and P4 in Fig. 7, which are compared against reference spectra measured with an spectrometer (Ocean Optics USB2000+). All the spectral curves are averaged in a 5×5 window, and normalized to the maximum value in their respective curves. The root-mean-square error (RMSE) between the reconstructed and the reference spectra are included in each subplot for ease of interpretation. It can be noticed that the curves from the use of RGB side information attain the best fit to the reference spectrum, followed by the curves from the reconstructions that use a grayscale side image, and these closely followed by the ones that use ToF, and the ones that do not use any, side information.

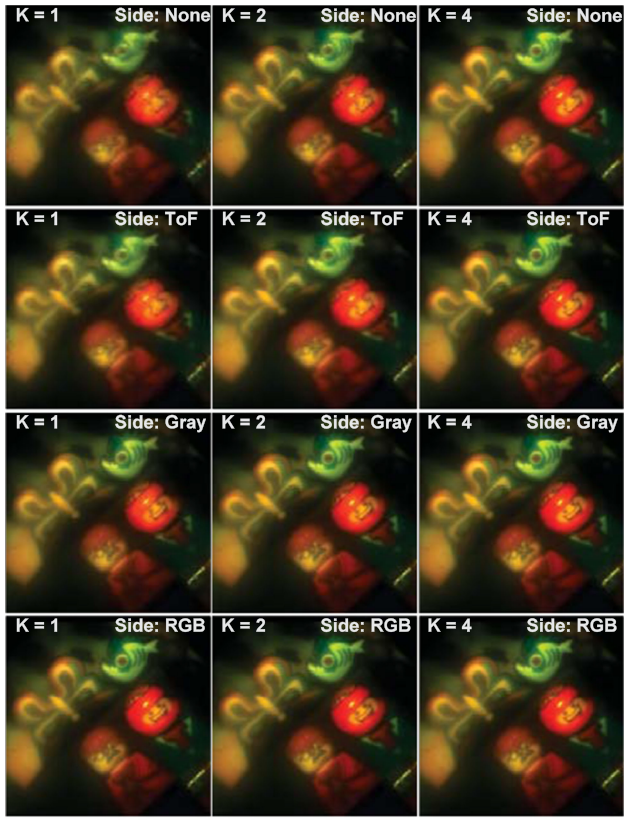
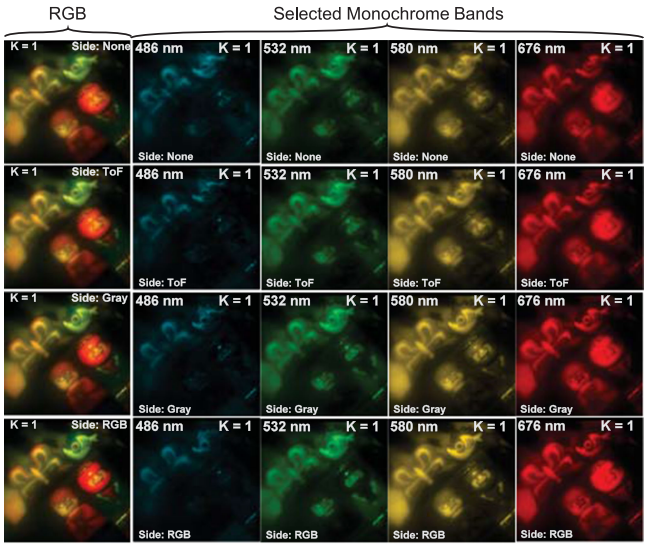
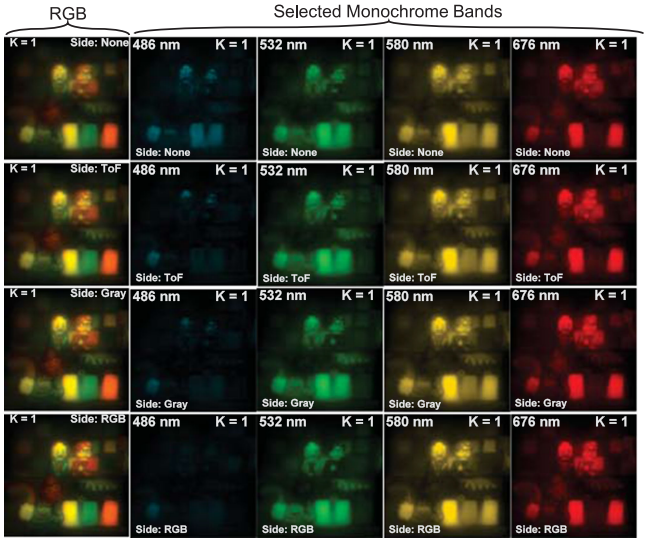


Fig. 10. Reconstructed data cubes mapped to RGB, when $K = 1, 2$ and 4 CSI snapshots are captured, using none, grayscale or RGB side information. (a) Scene 1. (b) Scene 2.



(a)



(b)

Fig. 11. Reconstruction results of 4 out of the 8 spectral bands, for a single snapshot, using none, grayscale or RGB side information. (a) Scene 1. (b) Scene 2.

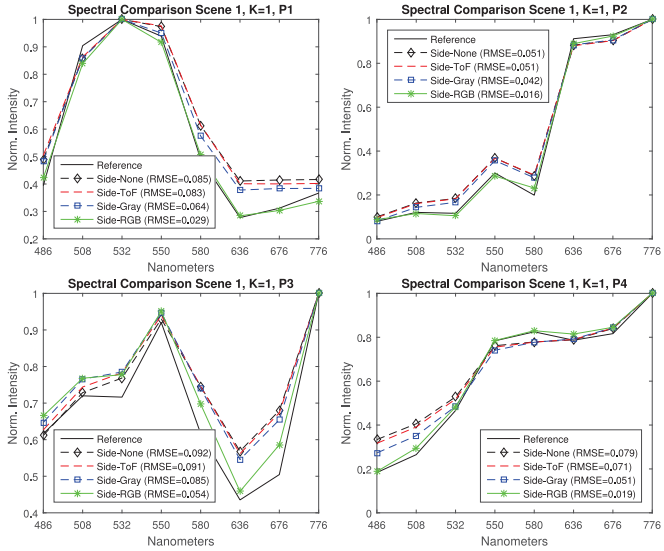


Fig. 12. Comparison of the spectrum reconstructions when $K = 1$ snapshots are captured with no side information, and with side information from ToF, grayscale or RGB camera. Four different points (P1 to P4) from the first target scene in Fig. 7(a) were measured by an spectrometer and compared against the reconstructed data cubes.

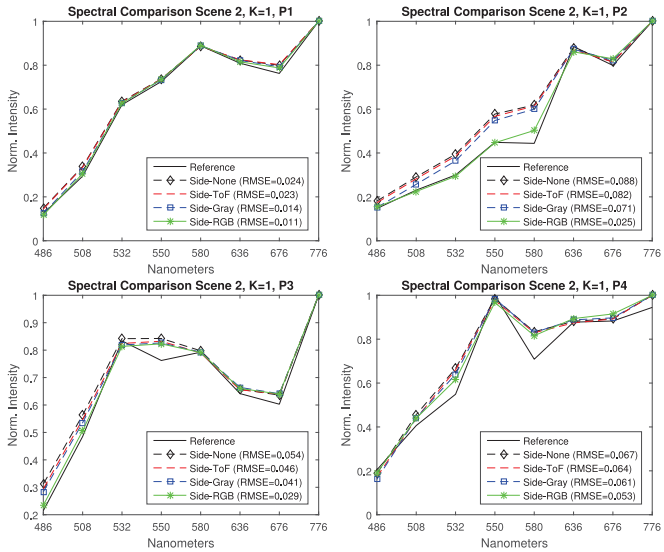


Fig. 13. Comparison of the spectrum reconstructions when $K = 1$ snapshots are captured with no side information, and with side information from ToF, grayscale or RGB camera. Four different points (P1 to P4) from the second target scene in Fig. 7(b) were measured by an spectrometer and compared against the reconstructed data cubes.

D. Reconstruction Results for ToF

To evaluate the 3D ranging estimation of the proposed testbed camera, the depth map is estimated using (10), and the sequence of 4 frames g_0 to g_3 captured in the ToF mode, shown in Fig. 9. Given the reduced range in which the toys are placed within the 2 scenes, we use different frequencies f for the modulation LED laser diode light, in order to evaluate the precision of the system. Recall that the target scenes are placed at around 40 and 50 cm away from the objective lens, respectively. Fig. 14 depicts the estimated depth maps using three modulation frequencies, $f = 6$ MHz, $f = 12$ MHz and $f = 24$ MHz, for the 2 target scenes, comparing the use of the CWMF described in Section II-C. Note that lower frequencies can reach longer ranges, but the depth precision will diminish. In Fig. 14(a) it can be seen that objects in the target scene 1 can be distinguished at 3 different depth planes, but $f = 24$ MHz attains a cleaner depth map than the other frequencies. Similarly in Fig. 14(b) the objects from the target scene 2 can be distinguished along 4 different depth planes, and again $f = 24$ MHz attains the most precise depth map. By comparing first against second column within Figs. 14(a) and 14(b), it can be noticed the boost in quality entailed by the use of the CWMF, since cleaner, and more precise, depth maps are achieved.

IV. Comparison With Alternative Solutions and Limitations of the Current System

Passive and active illumination 3D ranging systems have, indeed, their own pros and cons. Current off-the-shelf ToF sen-

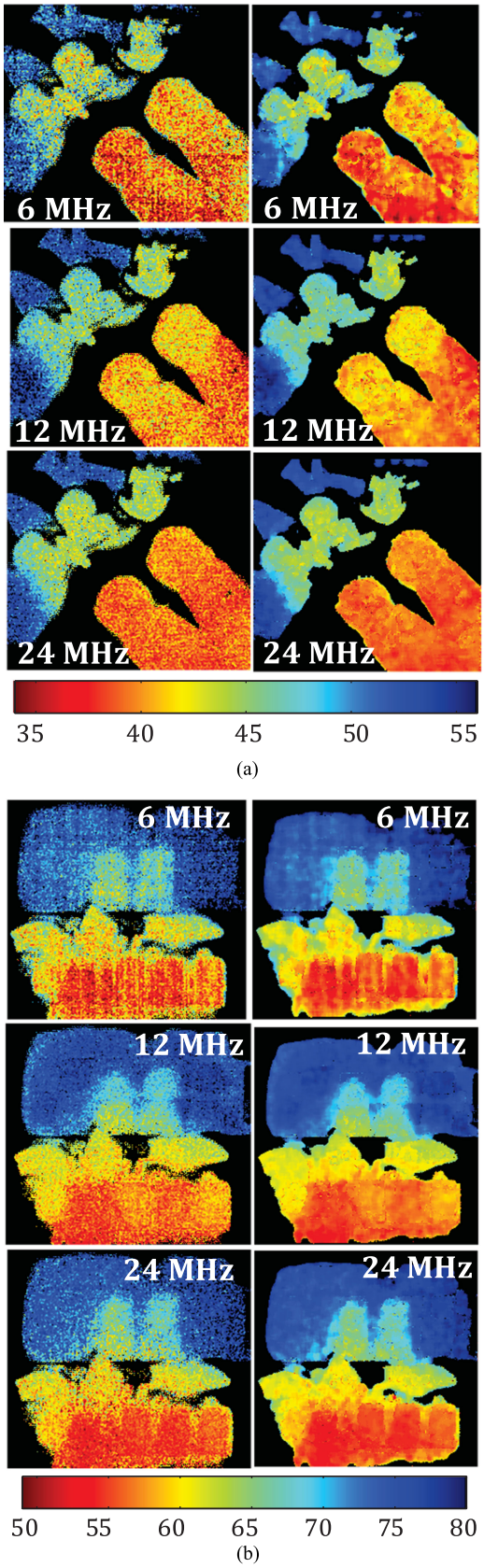


Fig. 14. Estimated depth maps using frequencies $f = 6$ MHz, $f = 12$ MHz, and $f = 24$ MHz. (a) Scene 1. (b) Scene 2. Units are in centimeters. (a) (Left) Depth from raw data. (Right) Depth using CWMF. (b) (Left) Depth from raw data. (Right) Depth using CWMF.

sors have low pixel count. However, these sensors are still in its early stages, and improving rapidly, as the technology becomes more popular for consumer applications. The proposed methods in this manuscript are general and will be applicable on ToF sensors as they continue to improve. Compared with stereo or structured light based depth cameras, ToF cameras exhibit several advantages. A big disadvantage of passive systems is that they cannot derive depth without depth cues created by texture; this is known as the depth stereo problem. Passive ranging methods, such as stereo imaging, and structured light imaging, rely on triangulation to interpolate depth values in the spaces between edges. The problem with triangulation is that it is very difficult to work over long ranges since the size of the triangle has to grow proportionally. They also have to deal with problems of occlusion where target surfaces are visible to the camera but not to the projector. ToF sensors, in contrast, do not suffer of occlusion problems, since the projector can wrap around the lens, thus the projector and camera have basically the same line of sight; and also do not require any texture or clue to estimate the depth. Moreover, ToF sensors can reach longer ranges by just changing the frequency and amplitude of the modulated light, they are lightweight, power efficient, and can collect full frame images with a single laser pulse, thus admitting very high frame-rates, and requiring no moving parts.

Regarding the single-aperture/single-sensor architecture, we expect that multi-mode (spectral + range) image sensors, such as the ToF used in our manuscript, will enable smaller form factors, lower power consumption and lighter weight imaging systems. This will lead to miniaturization and, in turn, to promote their use in commercial platforms.

While the benefits of the system in this paper are limited by the low pixel count of the ToF sensor, it does not, in reality, suffer from frame-rate reduction since the ambient light image and the modulated light amplitude can be jointly estimated from the same ToF quadrature components, using (7)–(13). Note however, that although we can estimate the 2 images from the quadrature components, we require at least 2 snapshots with complementary coded apertures, such that a pure (non-codified) modulated amplitude can be attained, in order to estimate the depth map per pixel. Note also that complementary snapshots will also boost the reconstruction quality of the spectral data cube estimation.

V. Conclusions and Future Work

A single aperture compressive snapshot multispectral+3D ToF ranging camera is proposed using a commodity ToF sensor, such as the Espros EPC660. The ambient light and modulated light dual-mode reading capability of the ToF sensor is exploited to jointly extract the characteristics of a target scene along the depth, space and spectral dimensions. It was shown that increasing the number of snapshots or by jointly reconstructing spectral and 3D ranging using side information, greatly improves the quality of reconstruction.

Given the pixel count of the current ToF sensor, multispectral+depth (MS+D) datacubes with 256×256 pixels of spatial resolution, 8 spectral channels and depth estimation at centime-

ter level, were recovered using compressive snapshots such that no scanning processes were required. The spatial resolution of the system can be further improved by using super-resolution techniques and high-resolution coded apertures [1], [44]. The spectral resolution can be enhanced by changing the dispersive element by a diffraction grating, so that hyperspectral+depth (HS+D) imagery could be attained. Further improvements along the spatial, spectral and depth dimensions can be achieved by optimizing the entries of the coded apertures [45], or by using fusion techniques that further exploit the side information.

REFERENCES

[1] G. R. Arce, D. J. Brady, L. Carin, H. Arguello, and D. S. Kittle, "An introduction to compressive coded aperture spectral imaging," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 105–115, Jan. 2014.

[2] D. Kittle, K. Choi, A. A. Wagadarikar, and D. J. Brady, "Multiframe image estimation for coded aperture snapshot spectral imagers," *Appl. Opt.*, vol. 49, pp. 6824–6833, 2010.

[3] H. Arguello, H. Rueda, Y. Wu, D. W. Prather, and G. R. Arce, "Higher-order computational model for coded aperture spectral imaging," *Appl. Opt.*, vol. 52, no. 10, pp. D12–D21, 2013.

[4] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1/3, pp. 7–42, 2002.

[5] I. Ihrke, J. Restrepo, and L. Mignard-Debise, "Principles of light field imaging: Briefly revisiting 25 years of research," *IEEE Signal Process. Mag.*, vol. 33, no. 5, pp. 59–69, Sep. 2016.

[6] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.

[7] J. Geng, "Structured-light 3D surface imaging: A tutorial," *Adv. Opt. Photon.*, vol. 3, pp. 128–160, 2011.

[8] J. Salvi, S. Fernandez, T. Pribanic, and X. Llado, "A state of the art in structured light patterns for surface profilometry," *Pattern Recognit.*, vol. 43, no. 8, pp. 2666–2680, 2010.

[9] S. Foix, G. Alenya, and C. Torras, "Lock-in time-of-flight (ToF) cameras: A survey," *IEEE Sensors J.*, vol. 11, no. 9, pp. 1917–1926, Sep. 2011.

[10] A. Bhandari and R. Raskar, "Signal processing for time-of-flight imaging sensors: An introduction to inverse problems in computational 3-D imaging," *IEEE Signal Process. Mag.*, vol. 33, no. 5, pp. 45–58, Sep. 2016.

[11] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from RGB-D images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 564–571.

[12] K. Lai, L. Bo, X. Ren, and D. Fox, "RGB-D object recognition: Features, algorithms, and a large scale benchmark, " in *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*. London, U.K.: Springer, pp. 167–192, 2013.

[13] P. Latorre-Carmona *et al.*, "Multispectral integral imaging acquisition and processing using a monochrome camera and a liquid crystal tunable filter," *Opt. Express*, vol. 20, no. 23, pp. 25960–25969, 2012.

[14] R. Horisaki, X. Xiao, T. Jun, and J. Bahram, "Feasibility study for compressive multi-dimensional integral imaging," *Opt. Express*, vol. 21, no. 4, pp. 4263–4279, 2013.

[15] I. Muniraj, B. Kim, and B. Lee, "Encryption and volumetric 3D object reconstruction using multispectral computational integral imaging," *Appl. Opt.*, vol. 53, pp. G25–G32, 2014.

[16] W. Feng, H. Rueda, C. Fu, Q. Chen, and G. R. Arce, "Compressive spectral integral imaging using a microlens array," *Proc. SPIE*, vol. 9857, 2016, Art. no. 985706.

[17] V. Farber, Y. Oiknine, I. August, and A. Stern, "Compressive 4D spectro-volumetric imaging," *Opt. Lett.*, vol. 41, no. 22, pp. 5174–5177, 2016.

[18] J. Wu, B. Xiong, X. Lin, J. He, J. Suo, and Q. Dai, "Snapshot hyperspectral volumetric microscopy," *Sci. Rep.*, vol. 6, 2016, Art. no. 24624.

[19] W. Feng, H. Rueda, C. Fu, G. R. Arce, W. He, and Q. Chen, "3D compressive spectral integral imaging," *Opt. Express*, vol. 24, no. 22, pp. 24859–24871, 2016.

[20] L. Wang, Z. Xiong, G. Shi, W. Zeng, and F. Wu, "Simultaneous depth and spectral imaging with a cross-modal stereo system," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.

[21] M. H. Kim *et al.*, "3D imaging spectroscopy for measuring hyperspectral patterns on solid objects," *ACM Trans. Graph.*, vol. 31, no. 4, 2012, Article no. 38.

[22] A. Kirmani, A. Colao, F. N. Wong, and V. K. Goyal, "Exploiting sparsity in time-of-flight range acquisition using a single time-resolved sensor," *Opt. Express*, vol. 19, no. 22, pp. 21485–21507, 2011.

[23] M. J. Sun *et al.*, "Single-pixel three-dimensional imaging with time-based depth resolution," *Nature Commun.*, vol. 7, 2016, Art. no. 12010.

[24] M. Van den Bergh and L. Van Gool, "Combining RGB and ToF cameras for real-time 3D hand gesture interaction," in *Proc. IEEE Workshop Appl. Comput. Vis.*, 2011, pp. 66–72.

[25] "Datasheet - epc660 3D TOF Imager 320 x 240 Pixel," Espros Photonics Corporation, [Online]. Available: ftp://ftp.espros.com/01_Chips/Datasheet_epc660-V1.08.pdf, Accessed on: Nov, 2016.

[26] S. Hsu, S. Acharya, A. Rafii, and R. New, "Performance of a Time-of-Flight range camera for intelligent vehicle safety applications," *Advanced Microsystems for Automotive Applications 2006*. New York, NY, USA: Springer, pp. 205–219, 2006.

[27] A. A. Wagadarikar, R. John, R. Willett, and D. Brady, "Single disperser design for coded aperture snapshot spectral imaging," *Appl. Opt.*, vol. 47, pp. B44–B51, 2008.

[28] H. Rueda, D. Lau, and G. R. Arce, "Multi-spectral compressive snapshot imaging using RGB image sensors," *Opt. Express*, vol. 23, pp. 12207–12221, 2015.

[29] C. Fu, M. Don, and G. R. Arce, "Compressive spectral imaging via polar coded aperture," *IEEE Trans. Comput. Imaging*, vol. 3, no. 3, pp. 408–420, Sep. 2017.

[30] G. R. Arce, H. Rueda, C. V. Correa, A. Ramirez, and H. Arguello, "Snapshot compressive multispectral cameras," *Wiley Encyclopedia of Electrical and Electronics Engineering*. Hoboken, NJ, USA: Wiley, 2017.

[31] N. Hagen, R. T. Kester, L. Gao, and T. S. Tkaczyk, "Snapshot advantage: A review of the light collection improvement for parallel high-dimensional measurement systems," *Opt. Eng.*, vol. 51, no. 11, pp. 111702:1–111702:7, 2012.

[32] N. Hagen and M. W. Kudenov, "Review of snapshot spectral imaging technologies," *Opt. Eng.*, vol. 52, no. 9, pp. 090901:1–090901:23, 2013.

[33] X. Yuan, P. Llull, D. J. Brady, and L. Carin, "Tree-structure Bayesian compressive sensing for video," arXiv: 1410.3080, 2014.

[34] M. F. Duarte and R. G. Baraniuk, "Kronecker compressive sensing," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 494–504, Feb. 2012.

[35] Y. August, C. Vachman, Y. Rivenson, and A. Stern, "Compressive hyperspectral imaging by random separable projections in both the spatial and the spectral domains," *Appl. Opt.*, vol. 52, pp. D46–D54, 2013.

[36] H. Rueda, H. Arguello, and G. R. Arce, "DMD-based implementation of patterned optical filter arrays for compressive spectral imaging," *J. Opt. Soc. Amer. A*, vol. 32, pp. 80–89, 2015.

[37] J. M. Bioucas-Dias and M. A. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2992–3004, Dec. 2007.

[38] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, Dec. 2007.

[39] A. Rajwade, D. Kittle, T. Tsai, D. Brady, and L. Carin, "Coded hyperspectral imaging and blind compressive sensing," *SIAM J. Imaging Sci.*, vol. 6, no. 2, 782–812, 2013.

[40] J. Tan, Y. Ma, H. Rueda, D. Baron, and G. R. Arce, "Compressive hyperspectral imaging via approximate message passing," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 2, pp. 389–401, 2016.

[41] L. Wang, Z. Xiong, D. Gao, G. Shi, and F. Wu, "Dual-camera design for coded aperture snapshot spectral imaging," *Appl. Opt.*, vol. 54, no. 4, pp. 848–858, 2015.

[42] X. Yuan, T. Tsai, R. Zhu, P. Llull, D. Brady, and L. Carin, "Compressive hyperspectral imaging with side information," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 964–976, Sep. 2015.

[43] K. E. Barner and G. R. Arce, *Nonlinear Signal and Image Processing: Theory, Methods, and Applications*. Boca Raton, FL, USA: CRC Press, 2003.

[44] M. L. Don, C. Fu, and G. R. Arce, "Compressive imaging via a rotating coded aperture," *Appl. Opt.*, vol. 56, pp. B142–B153, 2017.

[45] H. Arguello and G. R. Arce, "Rank minimization code aperture design for spectrally selective compressive imaging," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 941–954, Mar. 2013.



Hoover Rueda (S'12) received the B.Sc. and M.Sc. degrees in computer science from Universidad Industrial de Santander, Santander, Colombia, in 2009 and 2012, and the M.Sc. degree in electrical and computer engineering from the University of Delaware, DE, USA, in 2015. He is currently working toward the Ph.D. degree in electrical and computer engineering at the University of Delaware, sponsored by a Fulbright-Colciencias scholarship. His main research areas are computational optical imaging, compressed sensing, and high-dimensional signal processing.



Chen Fu (S'13) received the B.Sc. degree in electrical engineering from Nanjing University, Jiangsu, China, in 2012. He is currently working toward the Ph.D. degree in electrical and computer engineering at the University of Delaware under the supervision of Dr. G. R. Arce. His research interests include computational imaging, compressive sensing, optimization, and compressive spectral imaging.



Daniel L. Lau (SM'09) received the B.Sc. degree with highest distinction in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1995 and the Ph.D. degree from the University of Delaware, Newark, in 1999. He is currently the Kentucky Utilities Professor of Electrical and Computer Engineering, University of Kentucky, Lexington, and a Certified Professional Engineer. Prior to these appointments, he was a DSP Engineer at Aware, Inc., and an Image and Signal Processing Engineer at Lawrence Livermore National Laboratory. He is also Chief Technology Officer for Seikowave, Inc., a small Lexington company that sells 3-D scanners into the oil and gas pipeline services industry. Among his many published works is an article in the Proceedings of the IEEE and his own book, *Modern Digital Halftoning*, now in its second edition and published by CRC Press. His work has also been featured in such trade magazines as Vision Systems Design, Photonics Spectra, Imaging Insight, Prosilica Camera News, and Inspect Magazine.



Gonzalo R. Arce (F'00) received the M.Sc. and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA. He is the Charles Black Evans Distinguished Professor of Electrical and Computer Engineering at the University of Delaware, DE, USA. His research interests include computational imaging, coded aperture optimization, and compressive sensing research and development. He holds the J. P. Morgan-Chase Fellow Award and was awarded the 2010 and 2017 Fulbright-Nokia Distinguished Chair in Information and Communications Technologies. He has served as the Department Chair from 2000 to 2009. He was elected member of the Arkansas Academy of Engineering and received the DuPont Young Investigator Award and the NSF Research Initiation Award. He is a Fellow of the Center for Advanced Studies, University of Delaware. He has served as the Editor and the Guest Editor of several scientific journal publications of the OSA and SPIE. He holds 15 U.S. patents and is the author of four books in the areas of imaging and signal processing.

From the Calibration of a Light-Field Camera to Direct Plenoptic Odometry

Niclas Zeller, Franz Quint, and Uwe Stilla

***Abstract*—This paper presents a complete framework from the calibration of a plenoptic camera toward plenoptic camera based visual odometry. This is achieved by establishing the multiple view geometry for plenoptic cameras. Based on this novel multiple view geometry, a calibration approach is developed. The approach optimizes all intrinsic parameters of the plenoptic camera model, the 3D coordinates of the calibration points, and all camera poses in a single bundle adjustment. Our plenoptic camera based visual odometry algorithm, called direct plenoptic odometry (DPO), is a direct and semi-dense approach, which takes advantage of the full sensor resolution. DPO also relies on our multiple view geometry for plenoptic cameras. Tracking and mapping works directly on the micro images formed by the micro lens array and, therefore, has not to deal with aliasing effects in the spatial domain. The algorithm generates a semi-dense depth map based on correspondences between subsequent light-field frames, while taking differently focused micro images into account. Up to our knowledge, it is the first method that performs tracking and mapping for plenoptic cameras directly on the micro images. DPO outperforms state-of-the-art direct monocular simultaneous localization and mapping (SLAM) algorithms and can compete in accuracy with latest stereo SLAM approaches, while supplying much more detailed point clouds.**

***Index Terms*—Light-field, plenoptic camera calibration, plenoptic multiple view geometry, SLAM, visual odometry.**

I. Introduction

VISUAL Odometry (VO) and Simultaneous Localization and Mapping (SLAM) currently are some of the most addressed tasks in computer vision. Such approaches allow for tracking and mapping in unknown environments without any further infrastructure (e.g., GPS). In the last years VO and SLAM approaches for monocular and stereo cameras as well as active RGB-D sensors (e.g., structured light sensors) were driven forward. Nevertheless, while pure monocular approaches lack from scale awareness, stereo cameras or RGB-D sensors in general have large dimensions and therefore are impractical for certain applications.

During the last decades plenoptic cameras gained more and more interest [1]. While such cameras have dimensions similar to monocular cameras they are able to retrieve depth from single images based on the recorded light-field. Hence, plenoptic camera based VO would combine key features from both monocular and stereo approaches and closes the gap between the two.

In this article we present the complete workflow for plenoptic camera based VO. We introduce a new mathematical model for micro lens array (MLA) based light-field cameras and define a plenoptic multiple view geometry based on this model. This multiple view geometry leads us to a plenoptic camera calibration approach and builds the foundation for the Direct Plenoptic Odometry (DPO) algorithm. DPO combines advantages of monocular VO, like a single sensor system or scale invariance with static stereo and scale awareness at least for object distances in the range of a few meters. Our algorithm creates highly detailed, semi-dense point clouds (see Fig. 1) by working directly on pixel intensities in the micro images of a plenoptic camera.

A. Related Work

1) Plenoptic Camera Calibration: During the last years different methods were published to calibrate plenoptic cameras.

A method for correcting aberrations of the main lens based on the recorded 4D light-field inside the camera is presented in [2]. [3] presents a complete pipeline for the calibration of the MLA based on the specific case of a Lytro camera.

A complete mathematical model for unfocused plenoptic cameras is derived for the first time in [4]. To overcome the problem of feature point detection in the small micro images, [5] presents a calibration method which makes use of line features extracted from the micro images.

There exist several methods to perform calibration of focused plenoptic cameras [6]–[10]. All these methods model the projection from the object space to the virtual image space instead of projecting directly to the micro images. While [6] and [7] rely on a planar calibration target, [10] estimates the camera parameters based on a 3D object.

Manuscript received November 30, 2016; revised April 16, 2017 and June 22, 2017; accepted July 12, 2017. Date of publication August 11, 2017; date of current version October 23, 2017. This work was supported in part by the Baden-Württemberg-Stiftung in its program "Photonics, Microelectronics, Information Technology" and in part by the Federal Ministry of Education and Research of Germany in its program "FHprofUnt." The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Qing Wang. *(Corresponding author: Niclas Zeller.)*

N. Zeller is with the Faculty of Electrical Engineering and Information Technology, Karlsruhe University of Applied Sciences, 76133 Karlsruhe, Germany, and also with Technische Universität München, 80333 München, Germany (e-mail: niclas.zeller@hs-karlsruhe.de).

F. Quint is with the Faculty of Electrical Engineering and Information Technology, Karlsruhe University of Applied Sciences, 76133 Karlsruhe, Germany (e-mail: franz.quint@hs-karlsruhe.de).

U. Stilla is with the Department of Photogrammetry and Remote Sensing, Technische Universität München, 80333 München, Germany (e-mail: stilla@tum.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2737965

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

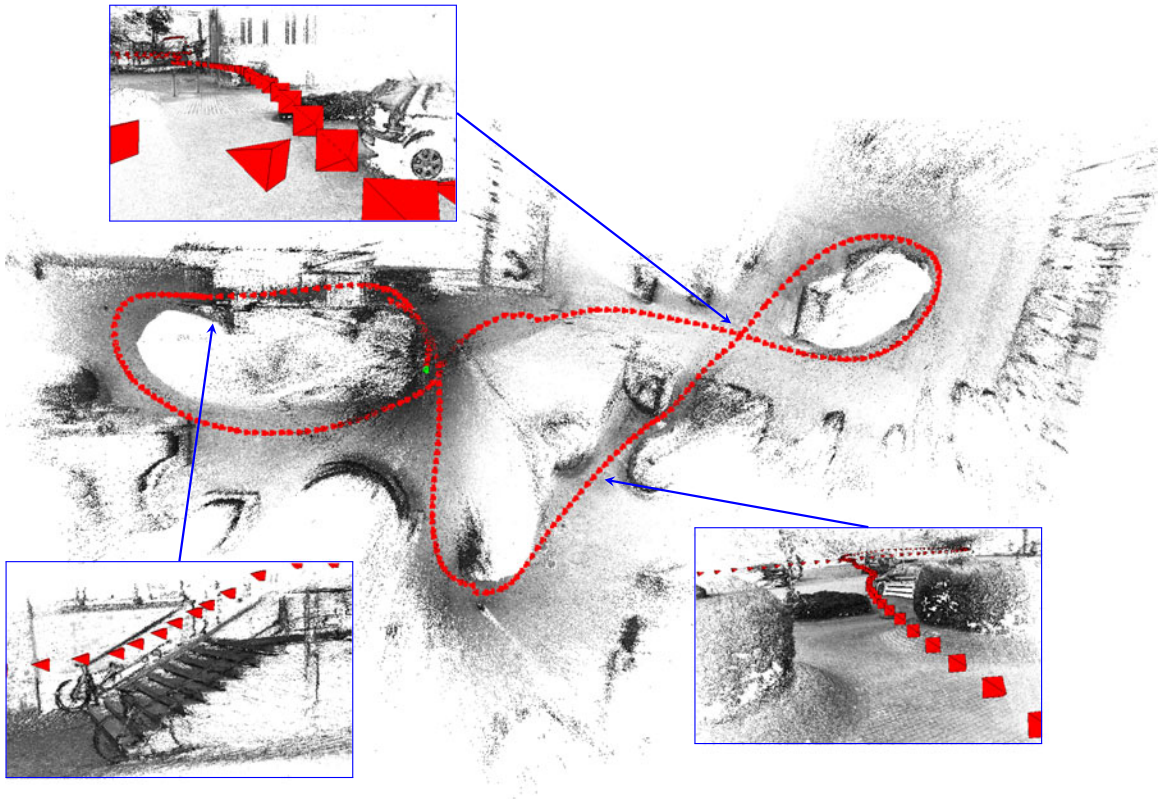


Fig. 1. Semi-dense point cloud of "parking lot" sequence created by Direct Plenoptic Odometry (DPO). The figure shows the top view of the complete trajectory and detailed subsections of the point cloud. Keyframe positions are marked in red, while the green dot represents the end of the trajectory.

2) Visual Odometry and SLAM: Feature based SLAM reduces the amount of data and therefore the computational effort by extracting a set of meaningful feature points from the recorded images. Afterwards, camera orientation and scene structure are estimated only based on this sparse set of geometric feature points [11]–[15].

Direct methods, like the one presented in [16], perform tracking and mapping directly on the recorded images. Tracking becomes much more robust since all image data is used. Here, tracking and dense mapping can be performed in one step, as shown in [17]. The complexity of such algorithms can be reduced by considering only textured image regions [18], [19]. These semi-dense, direct methods are capable to run in real-time on today's standard CPUs or even on smartphones [20].

While for monocular approaches camera motion is needed to obtain scene structure, the SLAM problem significantly simplifies when using stereo or RGB-D cameras. Since in this case absolute depth is received from a static recording, such approaches are able to measure the scale of the scene without using any additional sensors [21]–[24].

There exist few VO methods based on light-field representation [25], [26]. Anyhow, the cited methods mainly have been developed for light-fields recorded by camera arrays with large stereo baselines. In [27] a feature based Structure from Motion (SfM) approach working on 4D light-field representations is presented.

B. Contribution of This Work

This article presents a complete framework for plenoptic VO. We establish the multiple view geometry for MLA based light-field cameras. By working directly on the recorded micro images (raw image) we are able to find stereo correspondences using full sensor resolution instead of low resolution sub-aperture images. This avoids aliasing effects due to undersampling in the spatial domain [28] and results in much higher resolved depth maps.

Furthermore, we show that full resolution stereo matching leads to larger effective stereo baselines in comparison to the low resolution sub-aperture images.

We present a calibration approach which performs a complete bundle adjustment. Here, all intrinsic camera parameters, the 3D object coordinates and all camera poses are estimated in a single optimization task. The calibration is based on a 3D calibration target, which, as will be shown, significantly improves the calibration result compared to using a planar calibration target.

A major contribution is also the semi-dense direct visual odometry algorithm for MLA based light-field cameras, which makes use of the introduced multiple view geometry and the estimated camera model. We define a focus disparity error which models the effect of differently focused micro images on the depth estimation.

To the best of our knowledge, this is the first visual odometry approach which works directly on the recorded raw data of a

plenoptic camera and improves the static depth by plenoptic SfM.

C. Outline

In Section II we give some preliminary definitions used throughout this paper. Section III gives a brief, theoretical background on plenoptic cameras. The following methodical part consists of three main sections. Section IV defines the multiple view geometry for plenoptic cameras. The bundle adjustment based calibration approach is presented in Section V and Section VI describes the DPO algorithm. Finally, Section VII presents the evaluation of our methods, while Section VIII summarizes and concludes this work.

II. Preliminaries

Throughout the paper we use several notations, definitions and symbols. To enhance the readability of the paper, important notations, definitions and symbols are introduced here.

We denote matrices as bold, capital letters ($\boldsymbol{G}$), vectors as bold, lower case letters ($\boldsymbol{x}$) and scalars as normal letters, either capital or lower case (d). We do not differentiate between homogeneous ($\boldsymbol{x} = (x, y, z, 1)^T$) and non-homogeneous ($\boldsymbol{x} = (x, y, z)^T$) representations. Nevertheless, this should be clear from the context.

Throughout the paper rigid body transformations are defined based on Lie-Manifolds. Hence, a rigid body transformation $\boldsymbol{G} \in \mathrm{SE}(3)$ is completely defined by the six-dimensional vector $\boldsymbol{\xi} \in \mathbb{R}^6$, which defines an element of the corresponding Lie-Algebra $\mathfrak{se}(3)$. For more details on Lie-Manifolds we refer to [29].

The notation $[\boldsymbol{t}]_i$ defines the i-th element of a vector $\boldsymbol{t}$ and $[\boldsymbol{R}]_j$ the j-th row of a matrix $\boldsymbol{R}$.

The symbol δ_{mn} defines a Kronecker delta as given in eq. (1).

$$\delta_{mn} = \begin{cases} 1 & \text{for } m = n, \\ 0 & \text{for } m \neq n. \end{cases} \tag{1}$$

For the plenoptic camera we differentiate between image space and object space. To receive an upright image in image coordinates of an upright object in object coordinates, we define image coordinates to be mirrored with respect to object coordinates. Thus, to transform from object to image coordinates one has to take the negative coefficients of the respective vector.

Table I shows all important symbols used in this paper.

TABLE I
List of Important Symbols

symbol	description
v	virtual depth
b_L	image distance
$b_{L\,0}$	distance main lens to MLA
B	distance MLA to sensor
f_L	main lens focal length
f_M	micro lens focal length
D_M	micro lens diameter
$\boldsymbol{c}_L$	main lens principal point
$c_{L\,x}$	main lens principal point x-coord.
$c_{L\,y}$	main lens principal point y-coord.
$\boldsymbol{c}_{M\,L}$	micro lens center
$\boldsymbol{p}_{M\,L}$	projected micro lens center
$z_{C\,0}$	effective object distance offset
z_C	object distance
$z'_C = \lambda$	effective object distance
μ	pixel disparity
μ_p	projected micro image disparity
$\boldsymbol{x}_C$	camera coordinates
$\boldsymbol{x}'_C$	effective camera coordinates
$\boldsymbol{x}_R$	raw image coordinates
$\boldsymbol{x}_{M\,L}$	micro image coordinates
$\boldsymbol{x}_p$	projected micro image coordinates
$\boldsymbol{R}$	rotation matrix $\in \mathrm{SO}(3)$
$\boldsymbol{t}$	translation vector $\in \mathbb{R}^3$
$\boldsymbol{t}'$	effective translation vector $\in \mathbb{R}^3$
$\boldsymbol{\xi}$	element of the Lie-algebra $\mathfrak{se}(3)$
$\boldsymbol{G}$	rigid body transformation $\in \mathrm{SE}(3)$
$\boldsymbol{G}'$	effective rigid body trans. $\in \mathrm{SE}(3)$
d	inverse effective depth $d = \lambda^{-1}$
σ_d^2	inverse effective depth variance
$D_{M\,L}(\boldsymbol{x}_R)$	prob. micro image depth map
$D_V(\boldsymbol{x}_V)$	prob. virtual image depth map
$I_{M\,L}(\boldsymbol{x}_R)$	raw image (micro images)
$I_V(\boldsymbol{x}_V)$	virtual intensity image
$\sigma^2_{\mu(\xi,\pi)}$	variance of geometric disp. error
$\sigma^2_{\mu(I)}$	variance of photometric disp. error
$\sigma^2_{\mu(v,k,j)}$	variance of focus disparity error
$\pi_{M\,L}(\cdot)$	camera projection on micro images
$\pi_V(\cdot)$	camera projection on virtual image
$\delta_{m\,n}$	Kronecker delta

III. The Plenoptic Camera

Placing a MLA in front of the image sensor transforms a regular camera into a plenoptic camera, which is able to record 4D light-field information in a single image. While all plenoptic cameras rely on this concept, such cameras can be realized in different configurations depending on the camera parameters. In general we separate between unfocused plenoptic cameras (plenoptic camera 1.0) [30], [31] and focused plenoptic cameras (plenoptic camera 2.0) [32].

Focused plenoptic cameras produce focused micro images on the sensor, where each micro image captures a small portion of the complete scene. Due to the redundancy in the micro images, stereo correspondences can be found directly in the micro images using full sensor resolution. Hence, prior to any metric calibration one is able to estimate a so called *virtual depth* $v = \frac{b_L - b_{L\,0}}{B}$, which is a measure for the image distance b_L of the *virtual* main lens *image* [9] (see Fig. 2). Based on the virtual depth v one is able to project points from the focused micro images to the virtual image and thereby reconstruct the corresponding intensity image. Since this reconstructed image has a very large depth of field (DOF), we will call it the *totally focused image*.

From the recordings of unfocused plenoptic cameras one is able to extract a set of so called *sub-aperture images*. All of these sub-aperture images show the complete recorded scene from slightly different perspectives. The resolution of a sub-aperture image is limited by the number of micro lenses in the

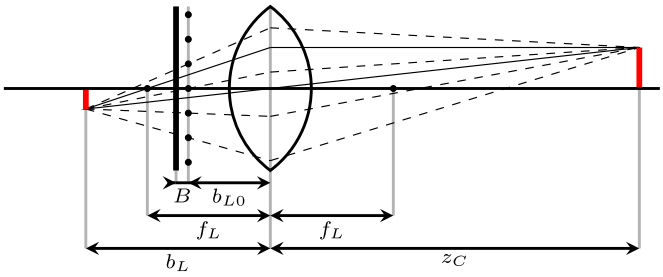


Fig. 2. Projection process of a focused plenoptic camera. MLA in distance b_{L0} to the main lens produces focused micro images of the virtual main lens image on the sensor.

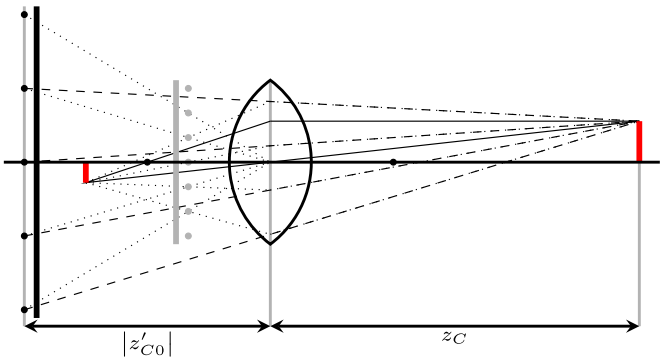


Fig. 3. Focused plenoptic camera interpreted as an array of very narrow FOV pinhole cameras at distance $|z'_{C0}|$ to the main lens.

MLA and therefore can not be increased by increasing the sensor resolution.

IV. Multiple View Stereo for Plenoptic Cameras

We show that a MLA based light-field camera can be interpreted as an array of very narrow field of view (FOV) pinhole cameras observing the scene. From this new interpretation insight into full resolution multiple view epipolar geometry for plenoptic cameras is obtained. Even though this interpretation is predestined for focused plenoptic cameras, such full resolution approaches can also be applied to unfocused plenoptic cameras, which supply partly focused micro images [33].

A. Plenoptic Camera Interpretation

Fig. 2 shows the projection process of a focused plenoptic camera. In the same way as the object points are projected through the main lens to corresponding image points, the micro lens centers can be projected from image space into object space based on the thin lens equation. Thus, the resulting object distance z'_{C0} of a micro lens center is defined as given in eq. (2).

$$z'_{C0} = \frac{f_L \cdot b_{L0}}{b_{L0} - f_L} \tag{2}$$

Here f_L is the focal length of the main lens and b_{L0} the distance between MLA and main lens plane. Fig. 3 shows the projected micro lens centers, resulting in a virtual camera array.

From Fig. 3 one can see that for the given setup ($f_L > b_{L0}$) the micro lenses are projected behind the main lens.

For later use we define z_{C0} as the negative value of z'_{C0}:

$$z_{C0} := -z'_{C0} = \frac{f_L \cdot b_{L0}}{f_L - b_{L0}} \tag{3}$$

With this definition, z'_C given in eq. (4) represents the object distance with respect to the virtual camera array.

$$z'_C := z_C + z_{C0} \tag{4}$$

The x- and y-coordinate of the center of a projected micro lens is received as the intersection between the projected MLA plane and the main lens' central ray through the corresponding real micro lens. Thus, one receives a projected micro lens center in camera coordinates $\boldsymbol{p}_{ML}$ from the coordinates of the real micro lens center $\boldsymbol{c}_{ML}$ as given in eq. (5).

$$\begin{aligned}\boldsymbol{p}_{ML} &= \begin{pmatrix} p_{MLx} \\ p_{MLy} \\ -z_{C0} \end{pmatrix} = -\boldsymbol{c}_{ML}\frac{z_{C0}}{b_{L0}} = -\begin{pmatrix} c_{MLx} \\ c_{MLy} \\ b_{L0} \end{pmatrix}\frac{z_{C0}}{b_{L0}} \\ &= -\boldsymbol{c}_{ML}\frac{f_L}{f_L - b_{L0}} = \boldsymbol{c}_{ML}\frac{f_L}{b_{L0} - f_L}\end{aligned} \tag{5}$$

Here, the minus in front of $\boldsymbol{c}_{ML}$ is due to the transformation from image coordinates to object coordinates.

The projected micro images are defined such that they have a normalized focal length and are parallel to the x-y-plane at $z_C = 1 - z_{C0}$. In this way the central projection from homogeneous 2D to 3D coordinates is just a scaling. A point $\boldsymbol{x}_p$ in the projected micro image is calculated based on a point $\boldsymbol{x}_{ML}$ in the real micro image as follows:

$$\begin{aligned}\boldsymbol{x}_p &= \begin{pmatrix} x_p \\ y_p \\ 1 \end{pmatrix} = \boldsymbol{x}_{ML}\frac{f_L - b_{L0}}{f_L \cdot B} + \frac{\boldsymbol{c}_{ML}}{f_L} \\ &= \begin{pmatrix} x_{ML} \\ y_{ML} \\ B \end{pmatrix}\frac{f_L - b_{L0}}{f_L \cdot B} + \frac{\boldsymbol{c}_{ML}}{f_L}\end{aligned} \tag{6}$$

Beside the regular camera coordinates $\boldsymbol{x}_C$ with its origin in the center of the main lens, we define for each micro lens, i.e., for each virtual camera, separate camera coordinates $\boldsymbol{x}'_C$. This is necessary since each micro lens has different center coordinates $\boldsymbol{p}_{ML}$ and thus the respective camera coordinate systems are also different. In the sequel we will call $\boldsymbol{x}'_C$ *effective camera coordinates*. The origin of the effective camera coordinates $\boldsymbol{x}'_C$ is the corresponding projected micro lens center $\boldsymbol{p}_{ML}$. Therefore, we receive the following relation:

$$\boldsymbol{x}_C := \boldsymbol{x}'_C + \boldsymbol{p}_{ML} = z'_C\boldsymbol{x}_p + \boldsymbol{p}_{ML} \tag{7}$$

Even though Figs. 2 and 3 show only a setup for $f_L > b_{L0}$ the presented projection is valid for almost any setup. The only setup for which no projection is possible is for $f_L = b_{L0}$. Here the micro lenses are projected to infinity. For $f_L < b_{L0}$ it could even be the case that the MLA is projected behind the recorded scene and effectively observes it from the backside.

B. Multiple View Epipolar Geometry

Based on the plenoptic camera interpretation (Section IV-A) we are able to define a multiple view epipolar geometry for plenoptic cameras.

Consider a set of images (views) taken with the plenoptic camera from different locations. Let i and j be two particular views of this set. Eq. (8) defines the transformation of a 3D object point with camera coordinates $\boldsymbol{x}_C^{(i)}$ in the i-th view to camera coordinates $\boldsymbol{x}_C^{(j)}$ in the j-th view based on the rigid body transformation $\boldsymbol{G}_{ij} \in \mathrm{SE}(3)$.

$$\boldsymbol{x}_C^{(j)} = \boldsymbol{G}_{ij} \cdot \boldsymbol{x}_C^{(i)} = \begin{pmatrix} \boldsymbol{R}_{ij} & \boldsymbol{t}_{ij} \\ \boldsymbol{0} & 1 \end{pmatrix} \cdot \boldsymbol{x}_C^{(i)} \tag{8}$$

Similarly to a point in regular camera coordinates, a point in effective camera coordinates $\boldsymbol{x}_C^{\prime(i,k)}$ of the k-th micro lens in the i-th view can be transformed into the effective camera coordinates $\boldsymbol{x}_C^{\prime(j,l)}$ of the l-th micro lens in the j-th view based on a rigid body transformation:

$$\boldsymbol{x}_C^{\prime(j,l)} = \boldsymbol{G}_{ij}^{\prime(kl)} \cdot \boldsymbol{x}_C^{\prime(i,k)} \tag{9}$$

with $\boldsymbol{G}_{ij}^{\prime(kl)} \in \mathrm{SE}(3)$ defined as follows:

$$\boldsymbol{G}_{ij}^{\prime(kl)} = \begin{pmatrix} \boldsymbol{R}_{ij} & \boldsymbol{t}_{ij}^{\prime(kl)} \\ \boldsymbol{0} & 1 \end{pmatrix} \tag{10}$$

$$\boldsymbol{t}_{ij}^{\prime(kl)} = \boldsymbol{t}_{ij} - \boldsymbol{p}_{ML}^{(l)} + \boldsymbol{R}_{ij} \cdot \boldsymbol{p}_{ML}^{(k)} \tag{11}$$

Here $\boldsymbol{p}_{ML}^{(k)}$ and $\boldsymbol{p}_{ML}^{(l)}$ are the coordinates of the respective projected micro lens centers as defined in eq. (5).

Based on this definition one can further derive the epipolar geometry between one micro image in the i-th view and another micro image in the j-th view. From projected image coordinates $\boldsymbol{x}_p^{(i,k)}$ (i-th view) and $\boldsymbol{x}_p^{(j,l)}$ (j-th view) one receives the effective camera coordinates $\boldsymbol{x}_C^{\prime(i,k)}$ and $\boldsymbol{x}_C^{\prime(j,l)}$ as follows:

$$\boldsymbol{x}_C^{\prime(i,k)} = \lambda_i \boldsymbol{x}_p^{(i,k)} \quad \text{with} \quad \lambda_i := z_C^{\prime(i,k)} \tag{12}$$

$$\boldsymbol{x}_C^{\prime(j,l)} = \lambda_j \boldsymbol{x}_p^{(j,l)} \quad \text{with} \quad \lambda_j := z_C^{\prime(j,l)} \tag{13}$$

In order to enhance readability we omit in the following all indices which do not lead to ambiguous definitions ($\boldsymbol{x}_C^{\prime(i)} := \boldsymbol{x}_C^{\prime(i,k)}$; $\boldsymbol{x}_C^{\prime(j)} := \boldsymbol{x}_C^{\prime(j,l)}$; $\boldsymbol{R} := \boldsymbol{R}_{ij}$; $\boldsymbol{t}' := \boldsymbol{t}_{ij}^{\prime(kl)}$).

Inserting eq. (12) and (13) into eq. (9) leads to the following relation:

$$\boldsymbol{x}_C^{\prime(j)} = \lambda_j \boldsymbol{x}_p^{(j)} = \lambda_i \boldsymbol{R} \cdot \boldsymbol{x}_p^{(i)} + \boldsymbol{t}' \tag{14}$$

where λ_j can be written as follows:

$$\lambda_j = \lambda_i \left[\boldsymbol{R}\right]_3 \cdot \boldsymbol{x}_p^{(i)} + \left[\boldsymbol{t}'\right]_3 \tag{15}$$

After combining eq. (14) and eq. (15) we receive a linear function in λ_i, as given in eq. (16).

$$\lambda_i \left(\left[\boldsymbol{R}\right]_3 \cdot \boldsymbol{x}_p^{(i)}\right) \boldsymbol{x}_p^{(j)} + \left[\boldsymbol{t}'\right]_3 \cdot \boldsymbol{x}_p^{(j)} = \lambda_i \boldsymbol{R} \cdot \boldsymbol{x}_p^{(i)} + \boldsymbol{t}' \tag{16}$$

Therefore, a point on the epipolar line in the micro image of the j-th frame is defined as follows:

$$\begin{aligned} \boldsymbol{x}_p^{(j)} &= \frac{\lambda_i \boldsymbol{R} \cdot \boldsymbol{x}_p^{(i)} + \boldsymbol{t}'}{\lambda_i \left[\boldsymbol{R}\right]_3 \cdot \boldsymbol{x}_p^{(i)} + \left[\boldsymbol{t}'\right]_3} \\ &= \frac{\lambda_i \boldsymbol{R} \cdot \boldsymbol{x}_p^{(i)} + \boldsymbol{t} - \boldsymbol{p}_{ML}^{(l)} + \boldsymbol{R} \cdot \boldsymbol{p}_{ML}^{(k)}}{\lambda_i \left[\boldsymbol{R}\right]_3 \cdot \boldsymbol{x}_p^{(i)} + \left[\boldsymbol{t}\right]_3 + z_{C0} + \left[\boldsymbol{R}\right]_3 \cdot \boldsymbol{p}_{ML}^{(k)}} \end{aligned} \tag{17}$$

Using this epipolar geometry we are able to deal with stereo observations from different micro images of different frames.

C. Effective Object Distances Versus Virtual Depth

For the case that the considered micro lenses are within the same frame ($\boldsymbol{t} = \boldsymbol{0}$ and $\boldsymbol{R} = (\delta_{mn})_{m,n \in \{1,2,3\}}$) eq. (17) simplifies as follows:

$$\boldsymbol{x}_p^{(j)} = \lambda_i^{-1} \left(\boldsymbol{p}_{ML}^{(k)} - \boldsymbol{p}_{ML}^{(l)}\right) + \boldsymbol{x}_p^{(i)} \tag{18}$$

From eq. (18) we receive the disparity in the projected micro images μ_p as given in eq. (19).

$$\mu_p := \frac{\langle \boldsymbol{x}_p^{(j)} - \boldsymbol{x}_p^{(i)}, \boldsymbol{p}_{ML}^{(k)} - \boldsymbol{p}_{ML}^{(l)} \rangle}{\|\boldsymbol{p}_{ML}^{(k)} - \boldsymbol{p}_{ML}^{(l)}\|} = \lambda_i^{-1} \|\boldsymbol{p}_{ML}^{(k)} - \boldsymbol{p}_{ML}^{(l)}\| \tag{19}$$

From eq. (19) we see that for in-frame depth estimation the disparity μ_p is proportional to the inverse effective object distance $\lambda^{-1} = z_C^{\prime -1}$. In another paper [34] we have shown that the disparity μ in the real micro images is proportional to the inverse virtual depth v^{-1}. Since we performed a linear mapping from μ to μ_p, this leads us to the conclusion that there has to be a linear relationship between $z_C^{\prime -1}$ and v^{-1}. This is proven in the following:

$$\begin{aligned} z_C' &= z_C + z_{C0} = \left(\frac{1}{f_L} - \frac{1}{b_L}\right)^{-1} + z_{C0} \\ &= \left(\frac{1}{f_L} - \frac{1}{v \cdot B + b_{L0}}\right)^{-1} + \frac{f_L \cdot b_{L0}}{f_L - b_{L0}} \end{aligned} \tag{20}$$

Rearranging eq. (20) gives the following definition of $z_C^{\prime -1}$ with respect to v^{-1}:

$$z_C^{\prime -1} = -\frac{\left(f_L - b_{L0}\right)^2}{B \cdot f_L^2} \cdot v^{-1} + \frac{f_L - b_{L0}}{f_L^2} \tag{21}$$

D. Effective Stereo Baseline—Unfocused Versus Focused

In Section IV-A we showed that a focused plenoptic camera can be interpreted as an array of very narrow FOV virtual cameras with high resolution (full resolution or plenoptic 2.0 rendering). An unfocused plenoptic camera can be interpreted as an array of wide FOV virtual cameras with low resolution [35] (plenoptic 1.0 rendering). Here the resolution of a sub-aperture image, observing the complete scene, is limited by the number of micro lenses in the MLA.

Even though the images of both concepts have quite different characteristics, the camera setups differ only by the focal length

of the micro lenses. For some cases the plenoptic 1.0 and 2.0 concept can even be applied to the same raw image [33].

In the following we compare the effective stereo baseline for both concepts and therefore the benefits with respect to 3D scene reconstruction.

For plenoptic 1.0 rendering the maximum stereo baseline $\Delta B_{1.0\,\max}$ results from the distance B between MLA and sensor, the micro lens diameter D_M, and the main lens focal length f_L [35]:

$$\Delta B_{1.0\,\max} = \frac{D_M \cdot f_L}{B} \tag{22}$$

Since the virtual camera array is formed at distance f_L in front of the main lens, the object distance z_C can be calculated based on the pixel disparity μ in the sub-aperture images as given in eq. (23).

$$z_C = \frac{\Delta B_{1.0\,\max} \cdot f_L}{D_M \cdot \mu} + f_L \tag{23}$$

Here μ is scaled by the micro lens diameter D_M which defines the size of a pixel in the sub-aperture image. From eq. (23) we receive the depth accuracy $\sigma_{z_C\,1.0}$ for plenoptic 1.0 rendering given in eq. (24).

$$\sigma_{z_C\,1.0} = \left|\frac{\partial z_C}{\partial \mu}\right| \cdot \sigma_\mu = \frac{(z_C - f_L)^2}{f_L^2} \cdot B \cdot \sigma_\mu \tag{24}$$

Using the plenoptic camera interpretation of Section IV-A we can calculate the effective stereo baseline and therefore the theoretical depth accuracy for plenoptic 2.0 (full resolution) rendering in a similar way. The effective stereo baseline $\Delta B_{2.0}(\kappa)$ for a pair of projected micro images is received as follows:

$$\begin{aligned}\Delta B_{2.0}(\kappa) &= \|\boldsymbol{p}_{ML}^{(k)} - \boldsymbol{p}_{ML}^{(l)}\| \\ &= \|\boldsymbol{c}_{ML}^{(k)} - \boldsymbol{c}_{ML}^{(l)}\| \frac{f_L}{b_{L0} - f_L} \\ &= \kappa \cdot D_M \cdot \frac{f_L}{b_{L0} - f_L}\end{aligned} \tag{25}$$

Here κ defines the multiple of D_M between the two micro lens centers. Therefore $\kappa \geq 1$ holds.

The object distance z_C is calculated based on the disparity μ_p in the projected micro images as given in eq. (26).

$$z_C = \frac{\Delta B_{2.0}(\kappa)}{\mu_p} - z_{C0} \tag{26}$$

Again, the depth accuracy $\sigma_{z_C\,2.0}$ is received from the standard deviation σ_μ of the pixel disparity as follows:

$$\begin{aligned}\sigma_{z_C\,2.0} &= \left|\frac{\partial z_C}{\partial \mu_p}\right| \cdot \sigma_{\mu_p} = \left|\frac{\partial z_C}{\partial \mu_p}\right| \cdot \left|\frac{\partial \mu_p}{\partial \mu}\right| \cdot \sigma_\mu \\ &= \frac{(z_C + z_{C0})^2 \cdot b_{L0}}{\kappa \cdot D_M \cdot z_{C0}} \cdot \frac{f_L - b_{L0}}{f_L \cdot B} \cdot s_{pixel} \cdot \sigma_\mu \\ &= \frac{(z_C + z_{C0})^2}{\kappa \cdot D_M} \cdot \frac{(f_L - b_{L0})^2}{f_L^2 \cdot B} \cdot s_{pixel} \cdot \sigma_\mu\end{aligned} \tag{27}$$

In eq. (27) $\left|\frac{\partial \mu_p}{\partial \mu}\right|$ defines the scaling from the pixel disparity μ to the disparity μ_p in the projected micro images. Here s_{pixel} defines the size of a pixel.

Using eq. (24) and eq. (27) we calculate the expected depth accuracies (standard deviations of the object distance z_C) for plenoptic 1.0 and 2.0 rendering. Fig. 4 shows the accuracies for both concepts, while using the same camera parameters ($f_L = 35$ mm, $B = 0.35$ mm, $b_{L0} = 34.3$ mm, $D_M = 0.1265$ mm, $s_{pixel} = 5.5$ μmm). For both rendering methods we chose $\sigma_\mu = 1$ pixel.

From eq. (24) and eq. (27) one can see that both curves are parabola shaped. Though, for the given setup $\sigma_{z_C\,1.0}$ has a much steeper slope than $\sigma_{z_C\,2.0}$ (see Fig. 4), which is due to a shorter effective stereo baseline.

For the shown setup ($b_{L0} < f_L$) the virtual camera array of the plenoptic 2.0 rendering lies behind the main lens, while the one for plenoptic 1.0 rendering lies always at distance f_L in front of the main lens. This leads to different minima in the curves and thus to an intersection of both curves, as can be seen from Fig. 4(b).

For smaller object distances the plenoptic 2.0 approach can use further apart micro lenses for stereo matching ($\kappa > 1$), which again leads to an improved accuracy.

Therefore, as can be seen from Fig. 4, the plenoptic 2.0 approach is always superior to the plenoptic 1.0 approach with respect to depth estimation. The κ-s shown in Fig. 4 are the first 10 received for a hexagonal arrangement of the MLA (1.00, 1.73, 2.00, 2.65, 3.00, 3.46, 3.61, 4.00, 4.36, 4.58).

We did not evaluate the functions for object distances closer than 0.5 m since here the images of both concepts can not be considered to be in focus anymore and thus the real accuracy will deviate from the curves. Anyway, in VO we are interested in larger object distances.

The only way to improve the depth accuracy of plenoptic 1.0 with respect to plenoptic 2.0 rendering is to reduce B. Though, this seems to be unfeasible due to the thickness of the MLA and the resulting impractical small F-number of the main lens (see F-number matching in [36]).

V. Plenoptic Camera Calibration

We present a plenoptic camera calibration approach which is based on the multiple view geometry introduced in Section IV-B. For calibration a complete bundle adjustment is performed, which optimizes the parameters of the plenoptic camera model, the 3D object coordinates and all camera poses at the same time. This is done based on recorded marker points, which are detected in the micro images of the camera.

Since we perform a complete bundle adjustment, we do not rely on any prior knowledge about the calibration target, like calibration points lying on a planar, regular checkerboard grid for instance. Furthermore, due to the 3D structure of our calibration target the optimization problem is better conditioned than those which are based on a planar target.

The calibration process determines the intrinsic parameters of the plenoptic camera, which are presented in Section IV-A, and additional distortion parameters. Our distortion model is

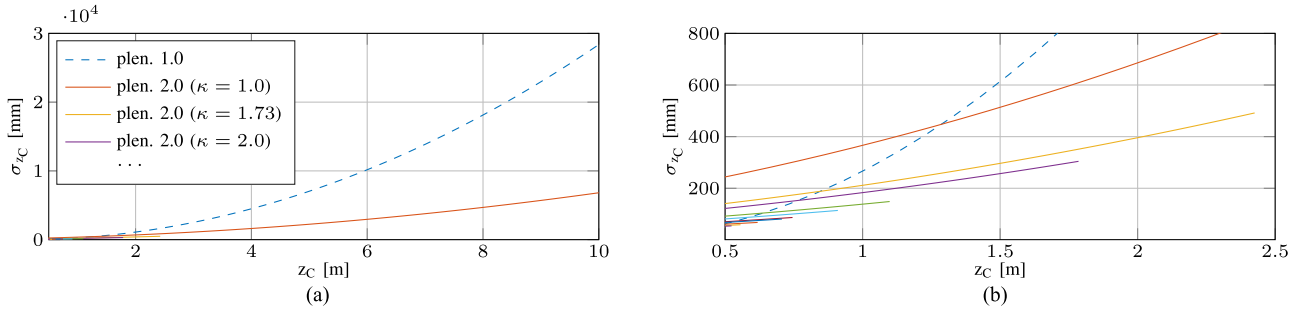


Fig. 4. Depth accuracy of plenoptic cameras based on plenoptic 1.0 and 2.0 rendering. Plenoptic 2.0 (full resolution) rendering results in a much larger stereo baseline and therefore better depth accuracy compared to plenoptic 1.0 approaches. (a) Depth accuracy. (b) Zoomed subsection of (a).

presented in Section V-A2. We do not consider the, in general marginal, effect of sensor tilting to preserve the planar arrangement of projected micro lens centers $\boldsymbol{p}_{ML}$ (as shown in Fig. 3). In Section V-B we present how to receive an initial solution for the bundle adjustment. The bundle adjustment, as we apply it to the calibration task, is presented in Section V-C.

A. Plenoptic Camera Model

1) Intrinsic Parameters: The plenoptic camera model consists of the following intrinsic parameters:

1) f_L – focal length of the main lens
2) b_{L0} – distance between MLA and main lens
3) B – distance between MLA and image sensor
4) $\boldsymbol{c}_L$ – principal point (intersection of the optical axis of the main lens with the image sensor)
5) $\boldsymbol{c}_{ML}^{(k)}$ – micro lens centers ($k \in \{1, 2, 3, \ldots\}$)

For the definition of these parameters we refer the reader to Section IV-B and Fig. 2.

To receive all intrinsic parameters in millimeters we define the pixel size, which we receive from the camera specification, as constraint. However, the correct pixel size is not relevant to obtain the optimum solution for the overall projection process.

While the micro lens centers $\boldsymbol{c}_{ML}^{(k)}$ are also part of the intrinsic camera parameters, we do not adjust them in the bundle adjustment. They can be easily estimated in advance using a white image (see [3] for instance). Furthermore, due to the small size of the micro images, in general we do not have more than one calibration point per micro image which would make the estimation of the micro lens centers in the bundle adjustment error prone.

2) Distortion Model: Most of the existing focused plenoptic camera models (i.e., [6], [7], [10]) define the projection from object space to virtual image space. Therefore, depth estimation has to be performed in advance using the distorted micro images and thus will be affected by the distortion of the main lens.

Since our model defines the projection from object space to the micro images, we are able to define the distortion model directly on the sensor before depth estimation. Hence, we do not have to deal with depth distortion afterwards.

Our distortion model considers radial and tangential distortion and is based on the well known model of Brown [37].

Here the radial distortion term Δr_{rad} is defined by a polynomial of the variable r, as given in eq. (28).

$$\Delta r_{rad} = A_0 r^3 + A_1 r^5 + A_2 r^7 + \cdots \tag{28}$$

The variable r is the distance between the principal point and the point coordinates on the MLA:

$$r = \sqrt{x^2 + y^2} \tag{29}$$

This results in the Cartesian correction terms Δx_{rad} and Δy_{rad} as given in eq. (30) and (31).

$$\Delta x_{rad} = x \frac{\Delta r_{rad}}{r} = x \left(A_0 r^2 + A_1 r^4 + A_2 r^6 + \cdots \right) \tag{30}$$

$$\Delta y_{rad} = y \frac{\Delta r_{rad}}{r} = y \left(A_0 r^2 + A_1 r^4 + A_2 r^6 + \cdots \right) \tag{31}$$

The tangential distortion terms (Δx_{tan} and Δy_{tan}) are defined as given in eq. (32) and (33).

$$\Delta x_{tan} = B_0 \left(r^2 + 2x^2 \right) + 2 B_1 xy \tag{32}$$

$$\Delta y_{tan} = B_1 \left(r^2 + 2y^2 \right) + 2 B_0 xy \tag{33}$$

Based on the correction terms the distorted coordinates $\tilde{x}$ and $\tilde{y}$ are calculated from the ideal projection as follows:

$$\tilde{x} = x + \Delta x_{rad} + \Delta x_{tan} \tag{34}$$

$$\tilde{y} = y + \Delta y_{rad} + \Delta y_{tan} \tag{35}$$

This model has the nice property that it consists only of additive distortion terms which depend on the undistorted coordinates. Due to the small size of a micro image it is sufficient to consider the distortion to be constant within a single micro image and therefore we only have to correct the micro image centers $\boldsymbol{c}_{ML}$ and not the point coordinates $\boldsymbol{x}_{ML}$ in the respective micro image.

By construction the micro lenses form a regular grid. The coordinates of their centers, which are estimated cf. [3] are the distorted micro lens centers $\tilde{\boldsymbol{c}}_{ML}$. Therefore, the micro lens centers $\boldsymbol{c}_{ML}$ which are corrected during the calibration considering the lens distortion and which are used for the projection, will deviate from this regular grid.

B. Initialization of the Calibration Process

While one is able to build a closed form solution for a SfM problem based on an uncalibrated monocular camera, this is not

the case for a plenoptic camera. The reason is that for each pair of micro images from distinct views one obtains a fundamental matrix. To estimate this one would need multiple corresponding points in each micro image, which, considering the small size of the micro images, is not realistic. This makes the initialization of the optimization problem more difficult. To initialize the bundle adjustment we consider in a first step that the totally focused image of the plenoptic camera results from a regular pinhole camera. In this way we solve the SfM problem using a standard photogrammetric software. For initialization this approximation is sufficient since only a rough estimate of all parameters is needed.

In addition, initial parameters for B and b_{L0} are received by solving the physical model defined in [9]. In this way we receive initial values for all intrinsic and extrinsic camera parameters as well as the 3D object points. A detailed description of the initialization can be found in [10].

Beside the initial parameters, the photogrammetric software provides already the correct correspondences of the recorded calibration points.

In the following we proof the validity of using the pinhole camera model for initialization, at least for $f_L \ll z_C$, based on a specific example. The camera setup is as follows:

1) main lens focal length $f_L = 16$ mm
2) size of the totally focused image: 1024×1024 pixel
3) principal point in the image center

The calibration target covers a depth range from $z_{C\,\min} = 500$ mm up to $z_{C\,\max} = 1500$ mm. Based on the thin lens equation we are able to calculate the following maximum, minimum and average image distances:

1) $b_{L\,\max} = 16.53$ mm
2) $b_{L\,\min} = 16.17$ mm
3) $\bar{b}_L = 16.35$ mm

Thus, one receives a maximum projection error between the plenoptic and pinhole camera model as follows:

$$\begin{aligned}\Delta x_{\max} &= x_{\max} \cdot \left(\frac{b_{L\,\max}}{\bar{b}_L} - 1 \right) \\ &= x_{\max} \cdot \left(1 - \frac{b_{L\,\min}}{\bar{b}_L} \right) = 5.63 \text{ pixel} \end{aligned} \tag{36}$$

Here the image plane of the pinhole camera model is considered to be in distance $\bar{b}_L$ to the main lens. The maximum image coordinate $x_{\max}$ is considered to be at the image boundary and therefore at $x_{\max} = 512$ pixel. As can be seen from eq. (36) the error made for the initialization is only in the range of a few pixels.

For larger object distances the error further decreases (i.e., for $z_{C\,\min} = 1500$ mm and $z_{C\,\max} = 2500$ mm it follows $\Delta x_{\max} = 1.11$ pixel). The variation in the average image distance $\bar{b}_L$ results in a small scaling error of the recorded image and therefore a bias in the estimated object distance. For $f_L \ll z_C$ it follows $b_L \approx f_L$ and therefore the estimated principal distance of the pinhole camera model gives a good initialization for the main lens focal length f_L of the plenoptic camera.

C. Plenoptic Bundle Adjustment

To estimate the exact camera parameters we formulate the following non-linear optimization problem:

$$E(\Pi, \Xi, P) = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{L} \left\| \boldsymbol{r}_{(i,j,k)} \right\|^2 \cdot \theta_{(i,j,k)} \tag{37}$$

Here, Π is the set of all camera parameters (intrinsic and distortion parameters), Ξ is the set of all camera poses ($\Xi := \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M\}$) and P the set of all object point coordinates ($P := \{\boldsymbol{x}_W^{(1)}, \ldots, \boldsymbol{x}_W^{(N)}\}$). Each calibration point in a micro image results in a 2D residual vector $\boldsymbol{r}_{(i,j,k)}$. $\theta_{(i,j,k)}$ is a masking function which is 1 if the i-th object point is visible in the k-th micro image of the j-th view and zero otherwise.

The residual vector for the calibration point with coordinates $\boldsymbol{x}_{ML}$, corresponding to the i-th object point which is seen in the k-th micro image of the j-th camera view, is defined as follows:

$$\boldsymbol{r}_{(i,j,k)} = \pi_{ML} \left(\boldsymbol{G}(\boldsymbol{\xi}_j) \boldsymbol{x}_W^{(i)}, \boldsymbol{c}_{ML}^{(k)}, \Pi \right) - \boldsymbol{x}_{ML} \tag{38}$$

The function $\pi_{ML}(\cdot)$ defines the projection from camera to micro image coordinates of the k-th micro image. In eq. (37) the sum over i is the sum over all object points, the sum over j is the sum over all camera poses and the sum over k is the sum over all micro images. To become robust against outliers, $\theta_{(i,j,k)}$ can also be extended with any robust loss function. The vector $\boldsymbol{\xi}_j \in \mathfrak{se}(3)$ defines the rigid body transformation $\boldsymbol{G}(\boldsymbol{\xi}_j) \in \mathrm{SE}(3)$ from world coordinates to camera coordinates of the j-th view.

The optimal parameters ($\hat{\Pi}, \hat{\Xi}, \hat{P}$) are the one which minimize the cost function $E(\Pi, \Xi, P)$:

$$\{\hat{\Pi}, \hat{\Xi}, \hat{P}\} = \arg\min E(\Pi, \Xi, P) \tag{39}$$

We solve this highly nonlinear optimization problem using the Levenberg-Marquardt algorithm. The initialization of Π, Ξ and P has to guarantee that the optimization starts in the convex region around the optimum solution of $E(\Pi, \Xi, P)$.

The scaling of the object is received based on known distances between certain object points. Those distances are used as additional constraints in the optimization problem.

Our implementation of the plenoptic camera calibration is based on the Ceres Solver Library [38].

VI. Direct Plenoptic Odometry

Our DPO algorithm performs direct image alignment and semi-dense mapping directly on the micro images recorded by a plenoptic camera. It makes use of the geometric camera interpretation defined in Section IV. The geometric camera interpretation relies on known intrinsic camera parameters, as well as distortion corrected images. Therefore, in advance to run DPO, the plenoptic camera has to be calibrated as described in Section V.

A. Overview

Algorithm 1 shows the complete workflow. We give here a short overview and focus on the single steps in the following subsections (Sections VI-B–VI-F). A new recorded light-field frame is tracked based on a keyframe. Keyframes

Algorithm 1: Direct Plenoptic Odometry.

Initialization:
currenKeyFrame $\leftarrow$ get first light-field frame in sequence $I_{ML0}(\boldsymbol{x}_R)$;
/* frame initially consists only of raw intensity image $I_{ML}(\boldsymbol{x}_R)$ */
$D_{ML}(\boldsymbol{x}_R) \leftarrow$ do In-Frame depth estimation on *currentKeyFrame*; /* see Sec. VI-C2 */
$D_V(\boldsymbol{x}_V) \leftarrow$ calculate central perspective depth map from $D_{ML}(\boldsymbol{x}_R)$;
$D_V(\boldsymbol{x}_V) \leftarrow$ regularize depth map $D_V(\boldsymbol{x}_V)$ and remove outliers; /* see Sec. VI-C4 */
$D_{ML}(\boldsymbol{x}_R) \leftarrow$ remove outliers in $D_{ML}(\boldsymbol{x}_R)$ which were detected in $D_V(\boldsymbol{x}_V)$;
$I_V(\boldsymbol{x}_V) \leftarrow$ calculate totally focused image for *currentKeyFrame* using $D_V(\boldsymbol{x}_V)$ and $I_{ML}(\boldsymbol{x}_R)$;
/* indices of current keyframe depth maps and intensity images are ignored since they exist only once */
for $j = 1 \rightarrow$ end of sequence **do**
  currentFrame $\leftarrow$ get next light-field frame in sequence $I_{MLj}(\boldsymbol{x}_R)$;
  Tracking:
  $\{\boldsymbol{\xi}_{kj},\ pointUsage,\ trackingResidual\} \leftarrow$ estimate pose from *currentKeyFrame* to *currentFrame* using $D_V(\boldsymbol{x}_V)$, $I_V(\boldsymbol{x}_V)$, $I_{MLj}(\boldsymbol{x}_R)$ and $\boldsymbol{\xi}_{k(j-1)}$; /* see Sec. VI-E */
  Depth Estimation:
  $D_{ML}(\boldsymbol{x}_R) \leftarrow$ estimate Inter-Frame depth from *currentKeyFrame* to *currentFrame* using $I_{ML}(\boldsymbol{x}_R)$, $D_{ML}(\boldsymbol{x}_R)$, $I_{MLj}(\boldsymbol{x}_R)$ and $\boldsymbol{\xi}_{kj}$; /* see Sec. VI-C3 */
  /* new depth estimates are merged into existing depth map $D_{ML}(\boldsymbol{x}_R)$ as described in Sec. VI-D */
  $D_V(\boldsymbol{x}_V) \leftarrow$ recalculate central perspective depth map from $D_{ML}(\boldsymbol{x}_R)$;
  $D_V(\boldsymbol{x}_V) \leftarrow$ regularize depth map $D_V(\boldsymbol{x}_V)$ and remove outliers; /* see Sec. VI-C4 */
  $D_{ML}(\boldsymbol{x}_R) \leftarrow$ remove outliers in $D_{ML}(\boldsymbol{x}_R)$ which were detected in $D_V(\boldsymbol{x}_V)$;
  $I_V(\boldsymbol{x}_V) \leftarrow$ recalculate totally focused image for *currentKeyFrame* using $D_V(\boldsymbol{x}_V)$ and $I_{ML}(\boldsymbol{x}_R)$;
  Keyframe Selection:
  newKeyframeScore $\leftarrow$ calculate new keyframe score using $\boldsymbol{\xi}_{kj}$ and *pointUsage*; /* see Sec. VI-F */
  if *newKeyframeScore* $>$ *keyframeSelectionThreshold* **then**
    newKeyFrame $\leftarrow$ *currentFrame*;
    $D_{ML\text{tmp}}(\boldsymbol{x}_R) \leftarrow$ estimate In-Frame depth on *newKeyFrame*; /* see Sec. VI-C2 */
    $D_{ML}(\boldsymbol{x}_R) \leftarrow$ propagate $D_{ML}(\boldsymbol{x}_R)$ from *currentKeyFrame* to *newKeyFrame* using $\boldsymbol{\xi}_{kj}$;
    $D_{ML}(\boldsymbol{x}_R) \leftarrow$ merge $D_{ML}(\boldsymbol{x}_R)$ with $D_{ML\text{tmp}}(\boldsymbol{x}_R)$;
    currentKeyFrame $\leftarrow$ *newKeyFrame*;
    $D_V(\boldsymbol{x}_V) \leftarrow$ recalculate central perspective depth map from $D_{ML}(\boldsymbol{x}_R)$;
    $D_V(\boldsymbol{x}_V) \leftarrow$ regularize depth map $D_V(\boldsymbol{x}_V)$ and remove outliers; /* see Sec. VI-C4 */
    $D_{ML}(\boldsymbol{x}_R) \leftarrow$ remove outliers in $D_{ML}(\boldsymbol{x}_R)$ which were detected in $D_V(\boldsymbol{x}_V)$;
    $I_V(\boldsymbol{x}_V) \leftarrow$ recalculate totally focused image for *currentKeyFrame* using $D_V(\boldsymbol{x}_V)$ and $I_{ML}(\boldsymbol{x}_R)$;
  end
end

are selected light-field frames in respect to which tracking and mapping of all recorded frames takes place. In addition to the light-field (raw image), for each keyframe two depth maps (micro image depth map and virtual image depth map) are established and a totally focused intensity image is synthesized. After successful tracking of a light-field frame the current keyframe is updated, i.e., the depth maps are refined based on the tracked frame and the totally focused image is recalculated.

During progress of the DPO algorithm, the pose of newly recorded frames differs more and more from the one of the initially established keyframe. Based on a score we check for each new frame whether the old keyframe is still valid or whether the new frame has to be set as a new keyframe. For a new keyframe, in-frame depth estimation is performed based on its recorded light-field. The in-frame depth map is merged with the one propagated from the last keyframe. Hence only the depth maps of the current keyframe will be kept.

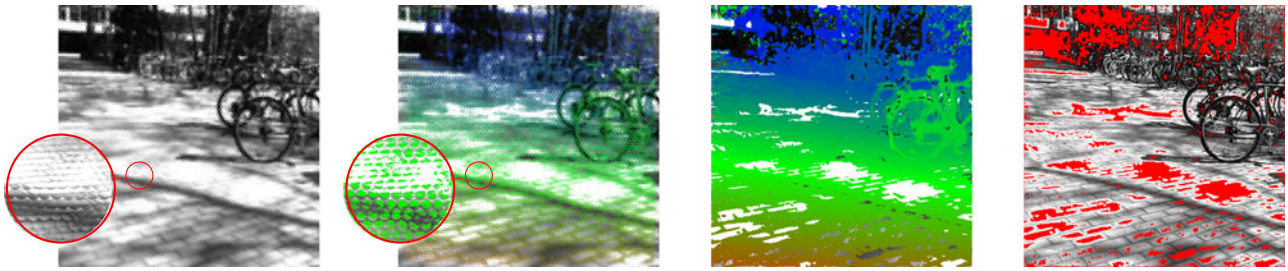
B. Depth Map Representation

For a plenoptic camera, the observed inverse virtual depth v^{-1} can be considers to result from a Gaussian process [34]. Therefore, due to eq. (21) the observed inverse effective depth $\lambda^{-1} = z_C'^{-1}$ has to result form a Gaussian process, too. Similar to [18] we define for each image point with an absolute gradient above a certain threshold a probabilistic depth hypothesis. We model the inverse effective depth as a Gaussian process:

$$\mathcal{N}\left(d, \sigma_d^2\right) \tag{40}$$

Here, $d = z_C'^{-1}$ defines the mean, while σ_d^2 is the corresponding variance.

1) Micro image Depth Map and Virtual Image Depth Map: Using the probabilistic depth model (eq. (40)) two different depth maps ($D_{ML}(\boldsymbol{x}_R)$ and $D_V(\boldsymbol{x}_V)$) are defined for each keyframe (Fig. 5(b) and (c)).



(a) (b) (c) (d)

Fig. 5. Keyframe depth maps and intensity images. (a) Micro images $I_{ML}(\boldsymbol{x}_R)$ (raw image recorded by the plenoptic camera). (b) Micro image depth map $D_{ML}(\boldsymbol{x}_R)$. (c) Virtual image depth map $D_V(\boldsymbol{x}_V)$. (d) Totally focused intensity image $I_V(\boldsymbol{x}_V)$. For the pixels marked in red no depth value and hence no intensity was calculated.

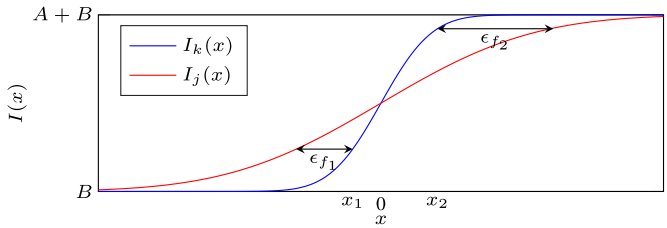


Fig. 6. Focus disparity error ϵ_f for two different positions (x_1 and x_2) on an edge along the epipolar line.

$D_{ML}(\boldsymbol{x}_R)$ is defined on raw image coordinates $\boldsymbol{x}_R$. Here a point in the i-th micro image, with micro image coordinates $\boldsymbol{x}_{ML}$ has the raw image coordinates $\boldsymbol{x}_R = \boldsymbol{x}_{ML} + \boldsymbol{c}_{ML}^{(i)}$. $D_V(\boldsymbol{x}_V)$ as well as a totally focused intensity image $I_V(\boldsymbol{x}_V)$ (Fig. 5(d)) are defined on virtual image coordinates $\boldsymbol{x}_V$ [34], where micro image points observing the same object point are merged (see Section VI-D).

We consider the two distinct depth maps because in the raw image one object point is projected to multiple image points, while there is a one to one mapping from object space to the virtual image.

Depth estimation is performed on $D_{ML}(\boldsymbol{x}_R)$, while the depth map $D_V(\boldsymbol{x}_V)$ is used for tracking new frames.

C. Estimating Depth Hypotheses

For each keyframe we perform in-frame depth estimation based on stereo matching in its own micro images as well as inter-frame depth estimation based on micro images of subsequent frames using the epipolar gemometry defined in Section IV-B. For both in-frame and inter-frame depth estimation stereo matches are found by optimizing the sum of squared intensity difference (SSID) over a one-dimensional pixel patch along the epipolar line. Prior estimates are used to narrow the search range to $d \pm 2\sigma_d$.

1) Observation Uncertainty: Similar to [18] we derive the observation uncertainty, which results in the variance σ_d^2 of the inverse effective depth d, based on uncertainty propagation. In addition to a geometric and a photometric disparity error, a focus disparity error, which regards unfocused micro images, is defined.

a) Geometric disparity error: Due to inaccuracies in the camera model and uncertainties in the frame pose one can expect a misalignment of the epipolar line. Due to the very short search range, we consider, similar to [18], only the absolute epipolar line position $\boldsymbol{l}_0$ to be effected by isotropic Gaussian noise ϵ_l and we can neglect any rotational error. Since for plenoptic cameras the search range is always limited by the micro lens dimension, the assumption of a short search range holds even better than for the monocular case. Thus, the variance $\sigma^2_{\mu(\xi,\pi)}$ of the geometric disparity error can be defined as follows [18]:

$$\sigma^2_{\mu(\xi,\pi)} = \frac{\sigma_l^2}{\langle \boldsymbol{g}, \boldsymbol{l} \rangle^2} \tag{41}$$

Here, $\boldsymbol{g}$ is the normalized image gradient, $\boldsymbol{l}$ the normalized epipolar line direction, and σ_l^2 the variance of ϵ_l.

b) Photometric disparity error: The photometric disparity error describes the effect of sensor noise ϵ_n on the estimated disparity. It results in an error on the estimated disparity with variance $\sigma^2_{\mu(I)}$, defined as follows [18]:

$$\sigma^2_{\mu(I)} = \frac{2 \cdot \sigma_n^2}{g_I^2} \tag{42}$$

Here σ_n^2 is the variance of ϵ_n and g_I the intensity gradient along the epipolar line.

c) Focus disparity error: Different to regular cameras, which in general are focused to infinity, the micro images of the plenoptic camera can not be considered to be in focus for the complete operating range, especially not for a multi-focus plenoptic camera [36]. Hence, the focusing itself also affects the stereo observation (Fig. 6). In the following we derive the variance $\sigma^2_{\mu(v,k,j)}$ of the focus disparity error.

Let k be the index of the micro image for which mapping is performed, while j is the one of the stereo reference. To model the focus disparity error we consider the real edge $I_{real}(x)$ which is observed in the micro images as a perfect Heaviside-step-function $h(x)$ with amplitude A and offset B along the epipolar line:

$$I_{real}(x) = A \cdot h(x) + B \tag{43}$$

The variable x is the position on the respective epipolar line relative to the step position μ_{0i} ($i \in \{k, j\}$):

$$x = \mu_k - \mu_{0k} = \mu_j \cdot \gamma - \mu_{0j} \tag{44}$$

Therefore, the correct disparity is defined by $\mu^* = \mu_{0j} - \mu_{0k}$. The parameters $\gamma = \frac{z_C'^{(j)}}{z_C'^{(k)}}$ defines the scaling factor between the micro images. During the imaging process the edge $I_{real}(x)$ is filtered by a Gaussian filter with variance σ_i^2 which depends on the virtual depth v, the micro lens type and the sampling rate (pixel pitch).

Thus, on the sensor we receive the following intensity functions along the epipolar line:

$$I_i(x) = B + \frac{A}{2}\left(1 + \operatorname{erf}\left(\frac{x}{\sigma_i\sqrt{2}}\right)\right) \quad i \in \{k, j\} \tag{45}$$

The estimated disparity $\hat{\mu} = \mu^* + \epsilon_f$ is the one for which both intensities have the same value and therefore, the following condition is fulfilled:

$$I_k(x) \stackrel{!}{=} I_j(x - \epsilon_f) \tag{46}$$

$$\operatorname{erf}\left(\frac{x}{\sigma_k\sqrt{2}}\right) \stackrel{!}{=} \operatorname{erf}\left(\frac{x - \epsilon_f}{\sigma_j\sqrt{2}}\right) \tag{47}$$

Since the error function ($\operatorname{erf}(\cdot)$) can not be solved analytically, we linearize eq. (47) and receive, after a rearrangement, the following relationship between the focus disparity error ϵ_f and the position on the edge x:

$$\epsilon_f = x \cdot \left(1 - \frac{\sigma_j}{\sigma_k}\right) \tag{48}$$

Fig. 6 visualizes the focus disparity error ϵ_f exemplary for two different positions (x_1 and x_2) on an edge along the epipolar line. Here, the red and blue curves represent the intensity along the epipolar line in two different micro images with different blur radii. For a certain position x on the edge the estimated disparity $\hat{\mu}$ will be the one for which both curves have the same value and therefore will be shifted by ϵ_f with respect to the real disparity μ^* as given in eq. (46). For the case that both images have the same blur radius, both curves are perfectly overlaid and therefore $\epsilon_f = 0$ will hold for any position x.

Considering the position with respect to the real edge x as a random variable with variance σ_x^2, the variance $\sigma_{\mu(v,k,j)}^2$ of the focus disparity error is as follows:

$$\sigma_{\mu(v,k,j)}^2 = \sigma_x^2 \cdot \left(1 - \frac{\sigma_j}{\sigma_k}\right)^2 \tag{49}$$

The standard deviation σ_i ($i \in \{k, j\}$) depends on the blur diameter s_i of the respective micro image, which is calculated based on the virtual depth v and the micro lens parameters (D_M, f_M, B) using the thin lens equation:

$$s_i = D_M \cdot \left|\frac{1}{v_i} + \frac{B}{f_{Mi}} - 1\right| \tag{50}$$

Since the minimum blur radius is limited by the pixel pitch and the overall optical system, the blur radius has a lower boundary s_0. Thus the following variances σ_k^2 and σ_j^2 result:

$$\sigma_k^2 = \beta^2 \cdot \min\{s_k^2, s_0^2\}, \quad \sigma_j^2 = \beta^2\gamma^2 \cdot \min\{s_j^2, s_0^2\} \tag{51}$$

The constant parameter β models the scaling from blur diameter to the standard deviation of the Gaussian filter.

Considering all three error sources as independent random variables, the overall observation uncertainty is received as follows:

$$\sigma_d^2 = \alpha^2 \cdot \left(\sigma_{\mu(\xi,\pi)}^2 + \sigma_{\mu(I)}^2 + \sigma_{\mu(v,k,j)}^2\right) \tag{52}$$

where α defines the derivative of d with respect to the disparity μ.

2) Estimating In-Frame Depth

In-frame depth estimation is performed similarly to the procedure we have described previously in [34]. The only difference is that here we directly estimate the inverse effective depth $z_C'^{-1}$ instead of the inverse virtual depth v^{-1}. Nevertheless, since there is a linear connection, the overall procedure stays the same.

3) Estimating Inter-frame Depth

While for in-frame depth estimation the stereo baseline is limited by the camera dimensions, we are able to improve the depth accuracy based on inter-frame depth observations. Inter-frame depth estimation is performed from the current keyframe to newly tracked frames. Therefore, we consider the relative orientation between the frames $\boldsymbol{\xi} \in \mathfrak{se}(3)$ to be known.

For each micro image point in the keyframe stereo observations are obtained from all possible micro images in the new frame.

a) Defining epipolar lines: One is able to define the epipolar geometry between the micro images of two different frames as given in Section IV-B. Due to the linear relation between projected micro image coordinates $\boldsymbol{x}_p$ and real micro image coordinates on the sensor $\boldsymbol{x}_{ML}$ (or $\boldsymbol{x}_R$), the epipolar lines defined in the projected micro images can be directly mapped on the sensor. Therefore, stereo matching for inter-frame depth estimation can be performed directly on the recorded raw images.

4) Regularizing Depth Maps

Each time the depth maps are updated a regularization step on $D_{ML}(\boldsymbol{x}_R)$ and $D_V(\boldsymbol{x}_V)$ is performed. Here, outliers are removed and estimates are smoothed based on probabilistic metrics, similar to [39]. Since it is much easier to detect outliers in $D_V(\boldsymbol{x}_V)$ the corresponding points in $D_{ML}(\boldsymbol{x}_R)$ are marked as outliers, too. The depth hypotheses in $D_{ML}(\boldsymbol{x}_R)$ itself are not updated based on $D_V(\boldsymbol{x}_V)$ to avoid loops in the filtering process. After updating $D_V(\boldsymbol{x}_R)$, the totally focused intensity image $I_V(\boldsymbol{x}_R)$ is recalculated, too.

D. Merging Depth Hypotheses

Probabilistic depth values are updated in a Kalman-like fashion, similar to [18]. However, in contrast to [18] we limit the variance of the merged depth hypothesis to the variance of the best observation. By this we do not consider the observations to be uncorrelated, which in the limit case would lead to zero variance. Thus, the following merging routine is defined:

$$\mathcal{N}\left(\frac{\sum_{i\in O} d_i \cdot \left(\sigma_{di}^2\right)^{-1}}{\sum_{i\in O}\left(\sigma_{di}^2\right)^{-1}}, \min\left(\sigma_{di}^2\right)\right) \tag{53}$$

where O is the set of depth observations. This algorithm avoids that points in the background which are observed in numerous frames receive low variances.

E. Tracking

Newly recorded frames are tracked based on direct image alignment. As tracking reference we use the current keyframe ($D_V(\boldsymbol{x}_V)$ and $I_V(\boldsymbol{x}_V)$).

Tracking is performed in two steps. Initial tracking is performed based on a coarse-to-fine approach using the totally focused image of the new frame. Final tracking is performed directly on the micro images.

1) Initial Tracking on Pyramid Levels

To obtain the totally focused image of the new frame without performing in-frame depth estimation, the virtual image depth map of the current keyframe is propagated to the new frame based on the pose of the last tracked frame. With this depth information the totally focused image can be synthesized.

For initial tracking we perform direct image alignment based on a coarse-to-fine approach as proposed in [40] to increases the radius of convergence. With the Levenberg-Marquardt algorithm we optimize the variance-normalized intensity error $E_V(\boldsymbol{\xi}_{kj})$:

$$E_V(\boldsymbol{\xi}_{kj}) = \sum_i \left\| \left(\frac{r_V^{(i)}}{\sigma_r^{(i)}} \right)^2 \right\|_\delta \tag{54}$$

$$r_V^{(i)} := I_{Vk}\left(\boldsymbol{x}_V^{(i)}\right) - I_{Vj}\left(\pi_V\left(\boldsymbol{G}(\boldsymbol{\xi}_{kj})\pi_V^{-1}(\boldsymbol{x}_V^{(i)})\right)\right) \tag{55}$$

Here $\pi_V(\cdot)$ defines the projection from camera to virtual image coordinates, while $\pi_V^{-1}(\cdot)$ is the inverse projection. $\|\cdot\|_\delta$ is the robust Huber norm [41].

2) Tracking on Micro Images

The result from the initial tracking is used to initialize the micro image based tracking. Here, each point in the tracking reference (frame k) $\boldsymbol{x}_C^{(i)}$ is projected to all micro images in the new frame (frame j) which observe this point. This results in the following optimization function which is solved with respect to $\boldsymbol{\xi}_{kj}$:

$$E(\boldsymbol{\xi}_{kj}) = \sum_i \sum_l \left\| \left(\frac{r_{ML}^{(i,l)}}{\sigma_r^{(i,l)}} \right)^2 \right\|_\delta \tag{56}$$

$$r_{ML}^{(i,l)} := I_{Vk}\left(\boldsymbol{x}_V^{(i)}\right) - I_{MLj}\left(\pi_{ML}\left(\boldsymbol{G}(\boldsymbol{\xi}_{kj})\pi_V^{-1}(\boldsymbol{x}_V^{(i)}), \boldsymbol{c}_{ML}^{(l)}\right)\right) \tag{57}$$

Here $\pi_{ML}(\cdot)$ defines the projection from camera to micro image coordinates. By the projection of the reference points to multiple micro images in the new frame we are able to implicitly incorporate the in-frame disparities of the new frame into the optimization problem. In this way we are able to avoid accumulated scale drifts as they in general occur for monocular tracking.

3) Variance of Tracking Residual

For both initial and final tracking the residual variance σ_r^2 is received based on uncertainty propagation. Therefore, σ_r^2 consists of a photometric component, which results from noise on the intensities and a geometric component which results from noise on the depth estimate.

$$\sigma_r^2 := \sigma_n^2 \left(\frac{1}{N_k} + \frac{1}{N_j} \right) + \left| \frac{\partial r(\boldsymbol{x}_V, \boldsymbol{\xi}_{kj})}{\partial d(\boldsymbol{x}_V)} \right|^2 \sigma_d^2(\boldsymbol{x}_V) \tag{58}$$

We consider the noise on the intensities in the different micro images to be uncorrelated. Therefore, the variance of an intensity value is estimated by the variance of the sensor noise σ_n^2 divided by the number of micro image points used to calculate the intensity value:

$$\left(\sigma_I^{(i)}\right)^2 = \frac{\sigma_n^2}{N_i} \quad \text{with} \quad i \in \{k, j\} \tag{59}$$

Here N_i ($i \in \{k, j\}$) is the number of micro image points used for calculation. Hence, for the tracking on the micro images it follows $N_j = 1$.

The geometric term is received based on uncertainty propagation from the variance σ_d^2 of the corresponding depth hypothesis.

F. Selecting Keyframes

A new keyframe has to be selected when the view of a new frame differs too much from the one of the current keyframe. Therefore, we define a score based on the translation between the two frames as well as on the percentage of good points during tracking.

When a frame is selected to be the new keyframe, in-frame depth estimation is performed in the new frame. Afterwards, the micro image depth map of the last keyframe is projected into the new one and the depth hypotheses are merged. Here, projected depths which are divergent from the in-frame depth are rejected and not merged.

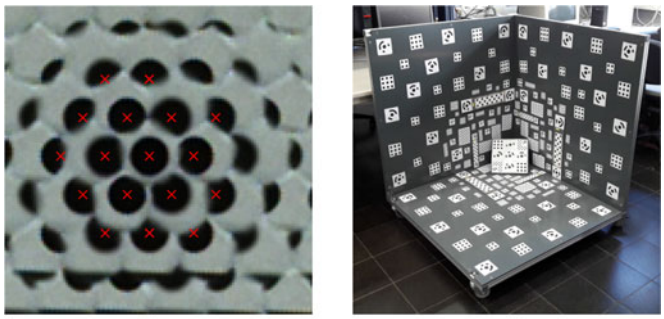
VII. Evaluation

A. Plenoptic Camera Calibration

In this section the evaluation of the calibration approach for the plenoptic camera is presented. We compare the calibration results for different main lenses. Besides, we compare our approach to a state-of-the-art calibration algorithm [5] for MLA based light-field cameras.

All experiments were performed with a Raytrix R5 camera (sensor size: 11,264 × 11,264 mm, image size: 2048 × 2048 pixel, diameter of micro lenses: ∼23 pixel, aperture: f/2.4).

Due to the reason that it is quite difficult to detect marker points directly in the micro images of the plenoptic camera, we detect them in our approach in the totally focused image and project them back to the micro images. Anyway, the totally focused image has to be calculated in advance to initialize the



(a) (b)

Fig. 7. 3D target used for plenoptic camera calibration. (a) Single marker point projected to multiple micro images. Detected marker coordinates are visualized by red crosses. (b) Complete 3D calibration target. Consists of (unique) coded and uncoded markers. Six defined distances between pairs of coded markers are used to retrieve the scale of the calibration object. (a) Marker point. (b) Calibration target.

TABLE II
ESTIMATED INTRINSIC CAMERA PARAMETERS AND REPROJECTION ERRORS FOR THREE DIFFERENT MAIN LENS FOCAL LENGTHS

Lens	f_L [mm]	b_{L0} [mm]	B [mm]	c_{Lx} [pixel]	c_{Ly} [pixel]	s_x [pixel]	s_y [pixel]
12.5 mm	12.616	11.789	0.353	1003.3	1043.9	0.213	0.220
16mm	16.273	15.482	0.357	1015.7	1056.3	0.123	0.126
35mm	34.868	33.993	0.367	1024.0	1042.4	0.063	0.067

bundle adjustment. Fig. 7(a) shows an example for the detected marker points in the micro images, while Fig. 7(b) shows the complete calibration target. In contrast, the method presented by Bok *et al.* [5] detects line features from a checkerboard pattern directly in the micro images.

1) Calibration With Different Main Lenses: We use three different main lenses with different focal lengths ($f_L = 12,5$ mm, $f_L = 16$ mm, $f_L = 35$ mm) to evaluate our calibration approach.

For each lens we estimate the camera parameters and we evaluate the reprojection error as well as the accuracy of the 3D points estimated during the bundle adjustment. This allows us to assess the validity of our model and the robustness of the calibration approach.

Table II shows the estimated intrinsic parameters for all three lenses and the corresponding reprojection errors. Here, s_x and s_y are the root mean square (RMS) values of the reprojection error in the x- and y-coordinate respectively.

From Table II one can see that for all three main lenses the reprojection errors are much smaller than one pixel. This confirms the validity of the complete projection model. Here it seems that the reprojection error is correlated with the main lens focal length f_L. Another indication for the validity of the defined projection model is that the estimated main lens focal length f_L is quite close to the nominal focal length of the respective lens. In addition, the estimated parameter B is similar for all three lenses. The parameter B is a constant of the plenoptic sensor and therefore does not depend on the main lens.

TABLE III
ACCURACY OF 3-D OBJECT COORDINATES ESTIMATED DURING THE BUNDLE ADJUSTMENT

Lens	12,5 mm	16 mm	35 mm
RMSE	0.289 mm	0.301 mm	0.332 mm
MAE	2.124 mm	2.352 mm	2.411 mm

TABLE IV
ROBUSTNESS OF THE ESTIMATED INTRINSIC CAMERA PARAMETERS USING A MAIN LENS FOCAL LENGTH OF $f_L = 16$ MM

Calibration Method		Our Method mean	st. dev.	rel. std.	Bok *et al.* [5] mean	std. dev.	rel. std [%]
f_L	[mm]	16.28	0.003	0.020%	–	–	–
b_{L0}	[mm]	15.53	0.018	0.118%	–	–	–
B	[mm]	0.36	0.007	1.975%	–	–	–
c_{Lx}	[pixel]	1015.07	0.705	0.069%	1000.09	13.651	1.365%
c_{Ly}	[pixel]	1046.89	0.390	0.037%	1048.73	11.051	1.054%
f_x	[pixel]	*2888.97*	*2.072*	*0.072%*	2963.39	6.406	0.216%
f_y	[pixel]	*2888.97*	*2.072*	*0.072%*	2963.01	6.577	0.222%
K_1		*2.04*	*0.007*	*0.319%*	1.93	0.094	4.907%
K_2		*689.49*	*14.917*	*2.164%*	767.41	20.988	2.735%

To evaluate the accuracy of the estimated 3D object coordinates, we register the point clouds received from the bundle adjustment for all three main lenses with a reference point cloud using ICP (iterative closest point). The reference point cloud is received based on SfM using a standard monocular camera. For this we used a professional photogrammetric measurement software to estimate a highly accurate reference point cloud. Table III shows the root mean square error (RMSE) and the maximum absolute error (MAE) between the point clouds received from the plenoptic camera based bundle adjustment and the reference point cloud. For all three main lenses we measure a RMSE of the point cloud of less than 1mm and thereby confirm the validity of the defined multiple view geometry. The MAE is less than 2.5 mm for all three lenses although also the outliers are considered.

A point cloud accuracy of less than 1mm is way less than the measurement accuracy of the plenoptic camera and therefore sufficient for a precise calibration of the camera.

2) Comparison to State of the Art: We compare our calibration approach to the one presented by Bok *et al.* [5]. Both methods define the complete projection from object space to the recorded micro images on the sensor. While the overall projection from object space to the micro images is quite similar for both models we mainly want to demonstrate that the robustness of the calibration significantly benefits from the 3D calibration target in comparison to the planar checkerboard used in [5].

To compare the methods, we recorded for both targets 64 images from different perspectives using the main lens with $f_L = 16$ mm.

To evaluate the robustness of the calibration approach we performed the calibration for each method 10 times by picking randomly a set of 20 images out of the complete collection of

TABLE V
TRACKING DRIFTS OVER THE COMPLETE SEQUENCE MEASURED BY LOOP CLOSURES

Sequence	ORB-SLAM2 [24] (stereo)			LSD-SLAM [19]			DPO		
	scale	rot. [°]	pos. [%]	scale	rot. [°]	pos. [%]	scale	rot. [°]	pos. [%]
park. lot	1.02^{-1}	10.44	1.96	12.56^{-1}	10.01	33.83	1.07^{-1}	1.55	0.33
foodcourt	1.01^{-1}	3.77	0.75	2.60^{-1}	4.73	9.17	1.08	2.20	0.96
office	–	–	–	1.36^{-1}	7.86	4.67	1.16^{-1}	3.31	1.15

The table shows scale factor, rotational drift, and absolute position error from the beginning to the end of the sequences. The position error is given in percentage of the sequence length.

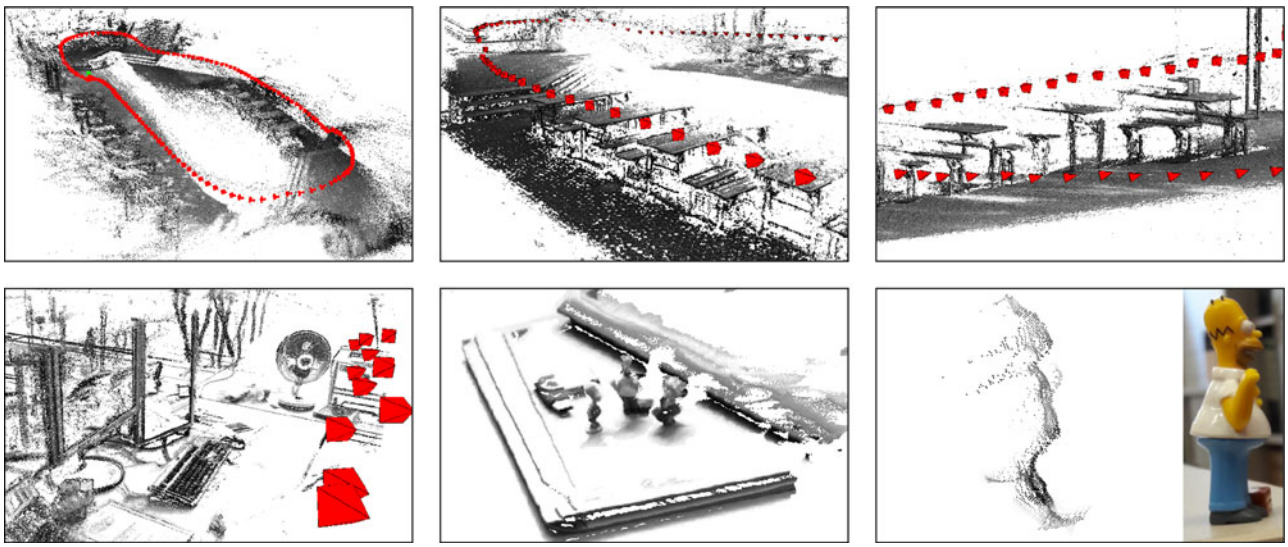


Fig. 8. Point cloud subsections created by DPO. Keyframe positions are marked in red. top: "foodcourt" sequence. bottom: "office" sequence. Homer figure (bottom, right) has a height of about 5cm.

64 images. The mean and the standard deviations of all intrinsic parameters are given for both methods in Table IV.

The parameters f_x, f_y, K_1 and K_2 need not to be calculated in our calibration approach. However, for comparison purposes we calculated these parameters (numbers in italic font) on the basis of our calibration results using the definitions given in [5].

As one can see from Table IV, the intrinsic parameters estimated with our calibration approach are an order of magnitude more robust than those of the method [5]. All estimates result in a relative standard deviation weigh less than one percent, except for the estimate of B. Probably the robustness of B could be improved by having more variation in the distance to the calibration target for the recorded images. A similar scattering as for B is received for K_2 since both parameters are highly correlated (see [5] for definition).

There is a somewhat large difference of the estimates of the intrinsic parameters f_x, f_y, K_1 and K_2 considering the two investigated methods. Our method estimates f_x and f_y to have the same value since we consider the pixels to be square. This is confirmed by the estimates of the method [5]. Since we do not have any absolute reference values we can not make any statement about the correctness of these values.

Both methods resulted in similar mean reprojection error, of 0.159 pixel for the method of Bok *et al.* [5] and 0.186 pixel for our method. This confirms the validity of both models.

B. Direct Plenoptic Odometry

Since there are no comparable plenoptic VO algorithms available we compare our method to state-of-the-art visual SLAM approaches: the monocular LSD-SLAM [19] and the stereo ORB-SLAM2 [24]. We ran our approach using the same camera as used for calibration (Section VII-A). The main focal length was chosen to $f_L = 16$ mm to achieve a suitable trade-off between FOV and scale awareness.

To compare the approaches, we mounted the plenoptic camera together with a stereo camera pair (image size: 1530×742 pixel, focal length: 705 pixel, baseline: 120 mm) on the same platform. We ran both camera systems with the same frame rate (15 fps) but without synchronization.

We present results for three different trajectories which were performed indoor as well as outdoor. For LSD-SLAM we reduced the image size to 640×480 pixel for which the implementation is optimized. For all sequences LSD-SLAM was initialized by the stereo depth map of the first frame, providing good starting conditions for this algorithm.

1) Quantitative Results: Since the frames of the sequences are not synchronized and we do not have a ground truth for the camera trajectory, we measure the overall drift based on a large loop closure. While we receive the loop closure data for the stereo camera sequences from the loop closure detection of

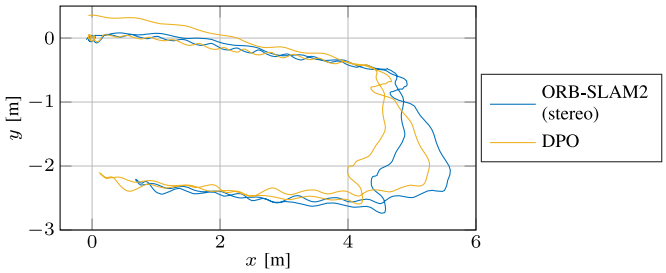


Fig. 9. Trajectory of "office" sequence estimated from ORB-SLAM2 (stereo) and DPO. The trajectory starts at coordinates (0,0,0), from there the cameras where moved on a U-shaped path around a group of tables and on a similar path back to the starting position.

ORB-SLAM2, we performed for DPO a 3D similarity transformation ($\mathrm{Sim}(3)$) based direct image alignment, as implemented in [19], between a representative keyframe at the beginning and the end of the sequence. The obtained co-registration from the beginning to the end of the sequences is used as ground truth data in the evaluation.

To obtain a fair comparison of all algorithms, for the evaluation large scale loop closure detection was disabled for ORB-SLAM2 and LSD-SLAM for all three sequences.

Table V presents the results for the three sequences. The table shows the scale factor, the rotational drift, and the absolute position error from the beginning to the end of the sequence.

For the "office" sequences ORB-SLAM2 results in a globally optimized trajectory since in the sequence a similar path was walked forward and backward (see Fig. 9). Therefore no drift could be measured. To avoid loop closures for LSD-SLAM on the "office" sequence we allowed map optimization only based on a neighborhood of ± 10 keyframes.

For the two large scale sequences ("parking lot" and "foodcourt") LSD-SLAM basically fails (scale drift $> 100\%$). Our DPO shows tracking accuracies comparable to ORB-SLAM2 (stereo) even though DPO performs only frame by frame tracking without any further optimization. DPO is only outperformed with respect to the scale drift, which is due to the much smaller static stereo baseline.

For all sequences ORB-SLAM2 (stereo) occasionally performed relocalization which means that tracking was lost temporarily and regained.

For the office sequence it might be that LSD-SLAM running on higher image resolution would result in less drifts. Nevertheless, for the two outdoor sequences ("parking lot" and "foodcourt") tracking of LSD-SLAM completely failed, which is not a result of the lower image resolution. The reason why LSD-SLAM fails is more because the camera is mainly moving towards viewing direction, which results in a biased $\mathrm{Sim}(3)$ estimation between consecutive keyframes and therefore in a continuously decreasing scale.

For both ORB-SLAM2 and LSD-SLAM we did not enforce real-time operation to obtain best possible results. Nevertheless, when rating the results given above, one has to consider that both algorithms are designed and implemented to run in real-time at the cost of poorer tracking and mapping performance. Our DPO in contrast currently exists only in a pure sequential, unoptimized implementation, which takes around 3 seconds to process a single frame. However, we showed already in previous publications [42] that depth estimation as a main part of DPO can run on frame rates > 30 fps on standard GPUs. Furthermore, direct image alignment on monocular images was also shown to be real-time capable [40]. Therefore, we are convinced that a real-time implementation of DPO, with similar performance as presented here, is feasible.

2) Qualitative Results: While on the outdoor trajectories ("parking lot" and "foodcourt") DPO has quite large absolute scale differences with respect to the stereo approach, on small scale it is able to estimate the scale from the light-field correctly. This is shown in Fig. 9, where the trajectories of DPO and ORB-SLAM2 (stereo) are shown for the "office" sequence.

Figs. 1 and 8 show subsections of the point clouds received from all three trajectories. Even though motion was performed mainly along the line of vision, DPO is still able to estimate accurate and detailed point clouds.

In close range DPO is able to estimate highly detailed point clouds with millimeter accuracy as can be seen from the silhouette of the Homer figure (Fig. 8, bottom, right), which was standing on the table in the "office" sequence and was recorded only from the front.

VIII. Summary and Conclusion

In this article we presented a complete framework to perform plenoptic camera based odometry. We developed a multiple view geometry for plenoptic cameras which enables tracking and mapping directly on the micro images at full sensor resolution. Based on this multiple view geometry we developed a calibration approach for plenoptic cameras which defines the projection from object space directly to the micro images on the sensor. By performing calibration based on a 3D calibration target, the resulting optimization problem is much better conditioned than in already existing methods. Our calibration method supplies a more robust estimate of the intrinsic camera model than a state-of-the-art calibration approach for MLA based light-field cameras.

The Direct Plenoptic Odometry (DPO) algorithm developed by us outperforms state-of-the-art monocular SLAM algorithms and is competitive to stereo approaches.

Although we do not have yet a real-time implementation of our DPO, we are convinced that an optimized implementation is able to run with high frame rates. Parts of the algorithm (depth estimation [42], direct image alignment [40]) have been shown to run at high frame rates.

Since there are no other plenoptic SLAM algorithms and respective datasets available yet, a direct comparison is not possible. Hence, this demands for a plenoptic odometry benchmark to compare future algorithms.

Acknowledgment

The authors would like to thank the anonymous reviewers and editors for their helpful comments.

REFERENCES

[1] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.

[2] R. Ng and P. Hanrahan, "Digital correction of lens aberrations in light field photography," *Proc. SPIE*, vol. 6342, 2006, Art. no. 63421E.

[3] D. Cho, M. Lee, S. Kim, and Y.-W. Tai, "Modeling the calibration pipeline of the Lytro camera for high quality light-field image reconstruction," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 3280–3287.

[4] D. Dansereau, O. Pizarro, and S. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 1027–1034.

[5] Y. Bok, H. G. Jeon, and I. S. Kweon, "Geometric calibration of micro-lens-based light field cameras using line features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 287–300, Feb. 2017.

[6] O. Johannsen, C. Heinze, B. Goldlücke, and C. Perwaß, "On the calibration of focused plenoptic cameras," in *Time-of-Flight and Depth Imaging: Sensors, Algorithms, and Applications* (ser. Lecture Notes in Computer Science). Berlin, Germany: Springer, 2013, pp. 302–317.

[7] C. Heinze, S. Spyropoulos, S. Hussmann, and C. Perwaß, "Automated robust metric calibration algorithm for multifocus plenoptic cameras," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 5, pp. 1197–1205, May 2016.

[8] N. Zeller, F. Quint, and U. Stilla, "Calibration and accuracy analysis of a focused plenoptic camera," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. II-3, pp. 205–212, Sep. 2014.

[9] N. Zeller, F. Quint, and U. Stilla, "Depth estimation and camera calibration of a focused plenoptic camera for visual odometry," *ISPRS J. Photogramm. Remote Sens.*, vol. 118, pp. 83–100, 2016.

[10] N. Zeller, C.-A. Noury, F. Quint, C. Teulière, U. Stilla, and M. Dhôme, "Metric calibration of a focused plenoptic camera based on a 3D calibration target," in *Proc. ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, 2016, vol. III-3, pp. 449–456.

[11] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. IEEE ACM Int. Symp. Mixed Augmented Reality*, Nov. 2007, vol. 6, pp. 225–234.

[12] E. Eade and T. Drummond, "Edge landmarks in monocular SLAM," *Image Vision Comput.*, vol. 27, no. 5, pp. 588–596, 2009.

[13] M. Li and A. I. Mourikis, "High-precision, consistent EKF-based visual-inertial odometry," *Int. J. Robot. Res.*, vol. 32, no. 6, pp. 690–711, 2013.

[14] A. Concha and J. Civera, "Using superpixels in monocular SLAM," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2014, pp. 365–372.

[15] R. Mur-Artal, J. M. M. Montiel, and J. D. Tards, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[16] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2014, pp. 15–22.

[17] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *Proc. IEEE Int. Conf. Comput. Vision*, 2011, pp. 2320–2327.

[18] J. Engel, J. Sturm, and D. Cremers, "Semi-dense visual odometry for a monocular camera," in *Proc. IEEE Int. Conf. Comput. Vision*, Dec. 2013, pp. 1449–1456.

[19] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 834–849.

[20] T. Schöps, J. Engel, and D. Cremers, "Semi-dense visual odometry for AR on a smartphone," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, Sep. 2014, pp. 145–150.

[21] S. Izadi *et al.*, "KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera," in *Proc. 24th Annu. ACM Symp. User Interface Softw. Technol.*, 2011, pp. 559–568.

[22] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 2100–2106.

[23] J. Engel, J. Stückler, and D. Cremers, "Large-scale direct SLAM with stereo cameras," in *Proc. Int. Conf. Intell. Robots Syst.*, Sep. 2015, pp. 1935–1942.

[24] R. Mur-Artal and J. D. Tards, "ORB-SLAM2: An open-source SLAM system for monocular, stereo and RGB-D cameras," *ArXiv preprint arXiv:1610.06475*, 2016.

[25] D. Dansereau, I. Mahon, O. Pizarro, and S. Williams, "Plenoptic flow: Closed-form visual odometry for light field cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 4455–4462.

[26] F. Dong, S.-H. Ieng, X. Savatier, R. Etienne-Cummings, and R. Benosman, "Plenoptic cameras in real-time robotics," *Int. J. Robot. Res.*, vol. 32, no. 2, pp. 206–217, Feb. 2013.

[27] O. Johannsen, A. Sulc, and B. Goldlücke, "On linear structure from motion for light field cameras," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 720–728.

[28] T. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 972–986, May 2012.

[29] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry, *An Invitation to 3-D Vision—From Images to Geometric Models* (ser. Interdisciplinary Applied Mathematics), 1st ed. New York, NY, USA: Springer, 2004.

[30] E. H. Adelson and J. Y. A. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 99–106, Feb. 1992.

[31] R. Ng, "Digital light field photography," Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, Jul. 2006.

[32] A. Lumsdaine and T. Georgiev, "Full resolution lightfield rendering," Adobe Systems, Inc., Mountain View, CA, USA, Tech. Rep. TR668, 2008.

[33] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *Proc. IEEE Int. Conf. Comput. Photography*, San Francisco, CA, USA, Apr. 2009, pp. 1–8.

[34] N. Zeller, F. Quint, and U. Stilla, "Establishing a probabilistic depth map from focused plenoptic cameras," in *Proc. Int. Conf. 3D Vision*, 2015, pp. 91–99.

[35] C. Hahne, A. Aggoun, S. Haxha, V. Velisavljevic, and J. C. J. Fernandez, "Baseline of virtual cameras acquired by a standard plenoptic camera setup," in *Proc. 3DTV-Conf.: The True Vision—Capture, Transmission Display 3D Video*, Jul. 2014, pp. 1–3.

[36] C. Perwaß and L. Wietzke, "Single lens 3D-camera with extended depth-of-field," *Proc. SPIE*, vol. 8291, Jan. 2012, Art. no. 829108.

[37] D. C. Brown, "Decentering distortion of lenses," *Photogramm. Eng.*, vol. 32, no. 3, pp. 444–462, May 1966.

[38] S. Agarwal *et al.*, "Ceres solver," 2012. [Online]. Available: http://ceres-solver.org

[39] N. Zeller, F. Quint, and U. Stilla, "Filtering probabilistic depth maps received from a focused plenoptic camera," in *Proc. 2nd BW-CAR Symp. Inf. Commun. Syst.*, 2015, vol. 2, pp. 7–12.

[40] C. Kerl, J. Sturm, and D. Cremers, "Robust odometry estimation for RGB-D cameras," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 3748–3754.

[41] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, Mar. 1964.

[42] R. Vasko, N. Zeller, F. Quint, and U. Stilla, " A real-time depth estimation approach for a focused plenoptic camera," in *Advances in Visual Computing* (ser. Lecture Notes in Computer Science), vol. 9475. New York, NY, USA: Springer, 2015, pp. 70–80.

Author's photographs and biographies not available at the time of publication.

Characterization of Plenoptic Imaging Systems and Efficient Volumetric Estimation From Plenoptic Data

Paul Anglin, Stanley J. Reeves, *Senior Member, IEEE*, and Brian S. Thurow

***Abstract*—Light-field cameras in conjunction with computational refocusing can be used to produce volumetric estimates of an imaged scene. However, these estimates are often dominated by image blur in the depth direction from objects not in each synthesized focal plane. Tomographic algorithms have been shown to be effective in creating volumetric estimates from plenoptic data but are often prohibitively slow. Deconvolution would be an attractive solution due to processing speed, but existing image synthesis equations are shift-variant. This research proposes an alternate refocusing transformation that makes the core problem described in continuous coordinates shift-invariant so that deconvolution is a viable solution. Shift-invariance of the new refocusing transform is demonstrated mathematically. Furthermore, the discretization involved in the imaging system and refocusing algorithm are characterized with respect to shift-variance in order to identify potential sources of artifacts and to propose potential mitigating steps where possible. While the sampled light field is not directly invertible, experimental data are used to demonstrate that regularized deconvolution using the derived synthesis equations produces improved results compared to the base focal stack in both synthetic examples and actual camera data.**

***Index Terms*—Deconvolution, plenoptic, light field.**

I. Introduction

TRADITIONAL imaging modalities such as a standard camera are limited to sampling 3D volumes as projections onto a 2D plane. These systems sample the spatial information present in the light field but lose the angular information. In order to retain the angular information inherent in the light field, movable or multi-camera systems can be used, but these are often limited by size and cost constraints [1]–[5]. However, light-field imaging offers a potential solution. A light-field—or plenoptic—camera samples each of the four dimensions of the light field with each exposure. By sampling both the spatial and angular domains, the user is given the tools to manipulate focus, perspective, and depth of field (DOF) during post-processing. With traditional imaging, these attributes are largely fixed once the image is acquired. The concepts related to sampling the 4D light field are not new and have been presented by many others [4], [6]–[8]. However, the works presented by [7], [9]–[11] and the introduction of the Lytro Camera have brought both these concepts and the technology to mainstream attention.

This work is concerned with estimating volumes from plenoptic image data. The estimation and reconstruction of volumetric processes and events from 2D measurements is of interest in a wide range of applications. Fluid flow measurement, estimation, and analysis via particle image velocimetry (PIV) is one application [1], [2], [12]–[24]. Volume reconstruction is also useful for analysis of flame structures, flame front propagation, and combustion events [3], [25]–[35]. 3D microscopy also requires volume reconstruction from 2D images [36], [37]. Many more applications exist, and improved understanding of these processes and events can lead to efficiency improvements and energy savings in the associated systems.

Using plenoptic data, a crude estimate of the imaged volume can be generated simply by creating computationally refocused images at slightly different depths throughout the volume. Algorithms for synthesizing images at varying focal depths from plenoptic data have been proposed [7], [9]–[11]. The resulting focal stack represents a kind of sampled version of the imaged volume. However, this representation is not ideal for all applications. It constitutes a poor estimate of the volume, as each refocused plane will contain energy from out-of-focus structures in adjacent focal planes. This out-of-focus energy is commonly recognized as image blur in conventional 2D images. In some scenarios, the image blur can dominate the response and obscure smaller structures in the image. In order to provide an accurate estimate of the volume, a method of reducing or eliminating the out-of-plane-energy is desired.

Tomographic methods utilizing plenoptic data have been developed to localize the energy in three dimensions. These methods have proven to produce largely blur-free estimates but are computationally expensive and may take hours to produce a single reconstruction [38]–[40]. In addition, the equations describing the relationship between the volume and the acquired data are relatively unstructured and do not lend themselves to a straightforward analysis of the problem. Other approaches have been developed from the perspective of deconvolution by identifying a point-spread function and then seeking to invert it [37], [41]. In the strict sense of the term, deconvolution refers

Manuscript received December 30, 2016; revised May 16, 2017 and August 25, 2017; accepted August 25, 2017. Date of publication August 30, 2017; date of current version October 23, 2017. This work was supported in part by the Air Force Office of Scientific Research, specifically under Grant FA9550-100100576 (Program Manager: Dr. D. Smith). The guest editor coordinating the review of this paper and approving it for publication was Prof. Yebin Liu. *(Corresponding author: Stanley J. Reeves.)*

P. Anglin and S. J. Reeves are with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849 USA (e-mail: jpa0005@auburn.edu; reevesj@auburn.edu).

B. S. Thurow is with the Department of Aerospace Engineering, Auburn University, Auburn, AL 36849 USA (e-mail: thurobs@auburn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2747481

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

to the inversion of a convolution process, which requires a shift-invariant system that is characterized by a single point-spread function (PSF). In [37], the system is acknowledged to be shift-variant, although a single PSF is used to characterize and remove the blur. A shift-invariant approximation may be somewhat justified in some applications of microscopy. The system is shift-invariant in the x and y directions in light-field microscopy; if the object is thin in depth, shift-variance in depth is minimal. For thicker objects, however, a shift-invariant approximation is not justified. In [41], the image formation is treated as shift-variant. Although the term deconvolution is used there, the algorithm employed is an iterative reconstruction algorithm that does not model the forward problem as shift-invariant. A shift-invariant form of the refocusing equation was used in [42] to develop a refocusing algorithm. However, a derivation and analysis of this form was not presented. Furthermore, other than our own preliminary work [43], we are not aware of previous papers that recognize or develop this shift-invariant mapping for volume reconstruction.

Volume reconstruction from plenoptic data in the general case is a large-scale problem that requires computationally intensive algorithms to solve. In this work we demonstrate that existing refocusing algorithms produce a shift-variant focal stack, and we reveal the primary nature of this shift-variance. Based on this insight we then define a modified mapping from data to optical image space and show that it yields a shift-invariant, continuous-space system that characterizes the mapping from the light source distribution in image space to the resulting refocused image stack. In the process we derive the PSF that characterizes the overall acquisition-refocusing system response. By exploiting this mapping we then solve the volume reconstruction problem via a standard frequency-domain deconvolution method that utilizes FFT-based processing techniques to convert matrix inversion into a pointwise division. The result is a computationally efficient algorithm that can produce volume reconstructions in seconds or minutes when given the focal stack and PSF. In addition to the computational advantages, the modified refocusing allows the plenoptic imaging problem to be interpreted more clearly in the frequency domain, leading potentially to greater insights into performance as well as future algorithmic developments.

We also present a detailed analysis of the characteristics of the combined acquisition-refocusing algorithms. This analysis reveals multiple sources of shift-variance not previously discussed, all of which arise from sampling the continuous-space system. We show that some of these characteristics can be mitigated with proper processing, while others are inherent to the system and cannot be improved upon.

To demonstrate an application of these shift-invariant equations, a basic deconvolution algorithm is used to reconstruct an experimental volume consisting of two separate flames. The volume is imaged using a plenoptic camera, and spatial-domain refocusing techniques are then used to synthesize the focal stacks from the experimental data. Finally, deconvolution is employed to reconstruct the imaged volume. Artifacts in the reconstructions are identified, and potential sources and methods to minimize them are addressed.

II. Background

The basic concepts related to light-field imaging, plenoptic cameras, and computational refocusing are presented here. However, the reader is encouraged to refer to works devoted to the subjects [4], [6]–[8], [10], [11], [44]–[48] for a fuller treatment of the topics. Helpful recent summaries can be found in [49], [50].

A. Plenoptic Function and Light-Field Imaging

The geometric distribution of light in the world around us can be described by a 5D function often referred to as the plenoptic function given by $P(x, y, z, \theta, \phi)$. Higher-dimensional variants exist, which can include time, wavelength, and polarization angle, but these variables are not considered in this work. This function has three spatial coordinates (x, y, z) and two angular coordinates (θ, ϕ) and is often used in computer vision and computer graphics. However, the measurement of light in a space is generally restricted to regions outside the convex hull, in which case the plenoptic function becomes redundant in one dimension. That is, the intensity of each ray does not change as it propagates in the z direction in the absence of occluders. In this case, the plenoptic function at a fixed depth z is sufficient to capture the entire plenoptic function. We refer to the resulting 4D function as the light field, which is denoted as $L(x, y, \theta, \phi)$ [4], [6].

Traditional imaging modalities capture a 2D projection of the 4D light field by integrating over (θ, ϕ). When the image is acquired by the camera sensor, only the (x, y) spatial information is recorded and the angular information is lost. However, a simple modification to a traditional camera allows the system to sample all four dimensions of the light field. By placing a microlens array between the camera main lens and the camera sensor, both the spatial and angular information of the imaged scene are encoded on the sensor. Placement of the microlens array with respect to the main lens focal plane determines how this information is encoded.

So-called Plenoptic 1.0 places the microlens array at the focal plane of the main lens. This subdivides the sensor such that the pixels underneath each microlens correspond to angular samples, and each microlens now corresponds to a spatial sample. This also establishes the fundamental tradeoff in Plenoptic 1.0 imaging. Subdividing the camera sensor in this way sacrifices spatial resolution for angular information. The system spatial resolution is now established by the number of lenses in the microlens array while the angular resolution is set by the number of pixels underneath each microlens [51], [52]. Such a tradeoff may have previously been unacceptable, but the explosive growth of sensor resolution has resulted in cases where the system resolution is limited by the optics rather than the sensor. For these cases as well as those where exceedingly high spatial resolution is not required, plenoptic imaging offers an attractive method of utilizing the otherwise unused sensor resolution [52].

Plenoptic 2.0—or focused plenoptic imaging—was later developed to address the loss of spatial resolution inherent in first-generation plenoptic imaging. Plenoptic 2.0 systems place and/or select the microlens array such that each microlens is

focused on the image plane formed by the camera main lens. This arrangement decouples the spatial resolution from the number of microlenses used in the microlens array at the expense of more computationally intensive and less intuitive processing algorithms. As a result, this work will focus on Plenoptic 1.0 to establish a solid foundation and understanding of the fundamentals associated with SI processing of light fields. Future work will consider more complex Plenoptic 2.0 processing techniques.

B. Computational Refocusing

Light-field cameras provide many post-processing opportunities that were not previously available to the photographer or researcher. The ability to computationally shift perspective and change the DOF are useful, but of particular interest to this work is the ability to create an image at any focal depth utilizing the data captured during a single acquisition. Computational refocusing is possible because a plenoptic camera captures not only the intensity at each spatial location but also the angle of arrival for the incident rays. With this information, the light field can be traced through space to simulate the equivalent sensor information at any arbitrary depth. The equivalent image at a chosen focal depth can be calculated by simply integrating the rays passing through each spatial location. In this way, images at any focal length are generated computationally in the same way that a traditional camera captures a single image, that is, by integrating over the angular information and retaining the spatial information.

C. Imaging Equations

A more rigorous description of the process can be achieved by referencing two significant planes in the camera [6]. The first plane corresponds to the camera main lens or aperture and is assigned coordinates (u, v) while the second plane corresponds to the microlens array and is assigned coordinates (x, y). This coordinate system is convenient, as any ray within the camera can now be described by its intersection with these two planes. Let $\boldsymbol{x} = [x\ y]'$ and $\boldsymbol{u} = [u\ v]'$. Then the irradiance at a given point on the $\boldsymbol{x}$ plane is given by

$$E_F(\boldsymbol{x}) = \frac{1}{F^2} \iint L_F(\boldsymbol{x}, \boldsymbol{u})\, d\boldsymbol{u} \tag{1}$$

where F is the distance between the $\boldsymbol{u}$ plane and the $\boldsymbol{x}$ plane. L_F is the light field given in terms of the two planes separated by F [9], [10]. This equation describes the formation of an image at the sensor plane of a standard camera. Determining the light field at the desired focal plane at distance $F-z$ from the aperture is accomplished by calculating the location where each ray intersects the new focal plane along the line:

$$\boldsymbol{x}_\alpha = \alpha \boldsymbol{x} - \boldsymbol{u}\,(\alpha - 1) \tag{2}$$

where $\alpha = \frac{F-z}{F}$ and $\boldsymbol{x}_\alpha$ is the coordinate in the new focal plane. This relationship is depicted in Fig. 1. Using this relationship and the fact that the light field is constant along these rays, the light field at $F-z$ is given by

$$L_{F-z}(\boldsymbol{x}_\alpha, \boldsymbol{u}) = L_F(\boldsymbol{x}, \boldsymbol{u}) = L_F\left(\boldsymbol{u}\left(1 - \frac{1}{\alpha}\right) + \frac{\boldsymbol{x}_\alpha}{\alpha}, \boldsymbol{u}\right) \tag{3}$$

These equations provide the necessary relationships for integral-based refocusing.

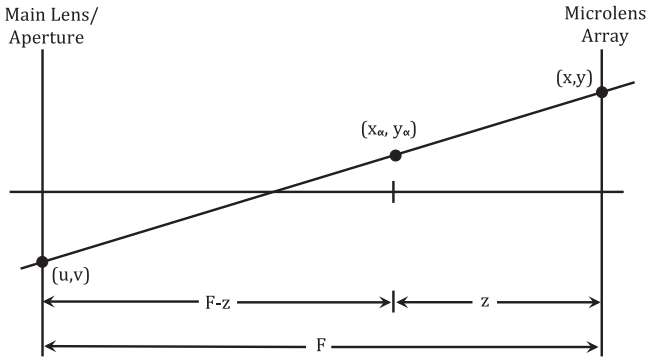


Fig. 1. An example light ray between a point (u, v) at the camera aperture and a point (x, y) at the microlens array is shown. A point along the ray at an arbitrary plane α is also depicted with the corresponding parameters annotated along the bottom of the figure.

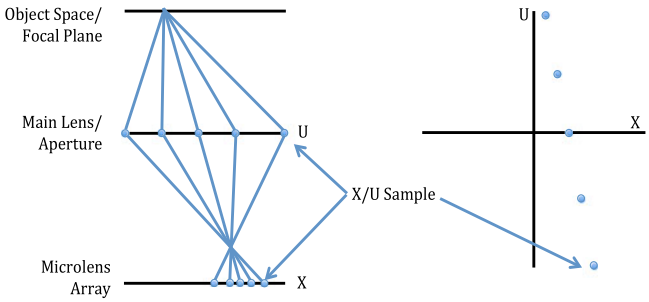


Fig. 2. The rays originating from an example point in object space and their path to the camera sensor are shown (left). The corresponding ray-space diagram is also shown (right). Each ray (left) between a given x and u corresponds to a single sample value at coordinate (x, u) in the ray-space diagram. The samples in the ray-space diagram are summed to produce a single point in the image.

D. Ray-Space Diagrams and the Radiance Array

The use of a two-plane parameterization also introduces the concept of a Cartesian ray-space diagram and the radiance array. These concepts have proven to be invaluable as a visualization tool in determining the source of spatial variability in the refocusing algorithm. Fig. 2 shows the relationship between the imaged point and the ray-space diagram in two dimensions. Light rays radiating from a discrete point in space trace a line through the (x, u) plane, where the angle of the line is dependent on the distance of the point from the focal plane. A continuous distribution of light is sampled spatially by the microlens array and angularly by the camera sensor. That is, the pixels underneath each microlens correspond to angular samples across the camera aperture. Each ray shown as a line in image space corresponds to a single sample of the plenoptic function, which is represented as a point in the ray-space diagram. Digital refocusing to each point in the plane involves summing the angular samples along

TABLE I
2-D SIMULATION PARAMETERS

Microlens Focal Length	f_l	0.500 mm
Microlens Pitch	p_l	0.125 mm
Pixel Pitch	p_p	0.0074 mm
Number of Pixels	n_p	1503
Sensor Size		10.5 mm
Number of Microlenses	n_l	89
Main Lens Focal Length	f_m	50 mm

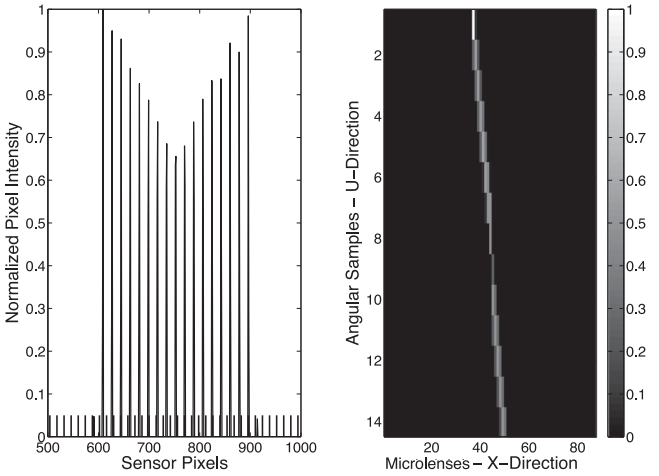


Fig. 3. Simulated sensor data with microlens locations indicated across the bottom (left) and the corresponding radiance array (right) for a point 9 mm in front of the optical focal plane.

a different line in the (x, u) plane, and the resulting sum is the synthesized pixel associated with the point in the plane. Moving the point in the x direction on the current focal plane simply shifts the line of summation along the x-axis in the ray-space diagram, while moving off the current focal plane results in a shearing or rotation of the line about the x-coordinate.

The physical application of the ray-space diagram can be seen when considering the sensor data acquired by a plenoptic camera. The camera sensor collects a 2D array of angular samples, but the spatial information is also encoded in the locations of these samples on the sensor. In order to simplify processing, it becomes beneficial to reorder the sensor data into a 4D radiance array that corresponds to the structure of the ray-space diagram.

To visualize and demonstrate concepts, 2D simulated data is presented throughout this work. This data is generated by choosing a point or points in object space and tracing the path of the light rays originating at these points back to the camera lens, to the microlens array, and finally to the sensor. The process is described in greater detail in [53], and the simulation parameters used are presented in Table I. Fig. 3 shows the relationship between the sensor data for a simulated point 9mm in front of the focal plane and the corresponding radiance array. Note the shearing/rotation of the ray-space samples that occurs for this point which is not on the optical focal plane. With this representation, refocusing to an arbitrary focal plane becomes a matter of choosing the correct angular samples from the radiance array and summing the values. This mapping will be exploited in subsequent sections.

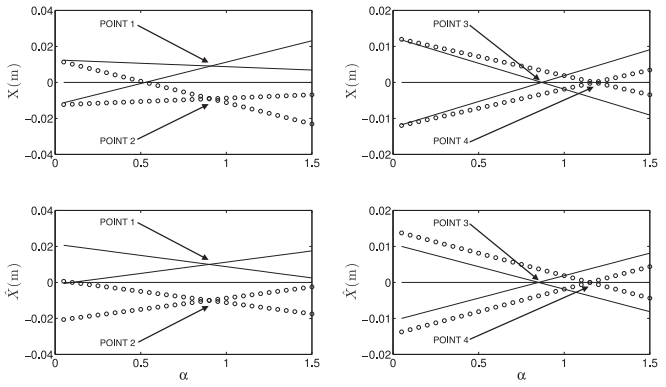


Fig. 4. The light path between the camera main lens and the microlense array for two example points is shown in each subfigure. The border of the light path for one point in each subfigure is shown with solid lines (–) while the border of the light path for a second point is shown with circles (○). The upper subfigures show shift-variant PSFs (top left and top right) showing spatial variability in x (top left) and in z (top right) – note the shape of the PSFs is different between the two points. The lower subfigures show shift-invariant PSFs (bottom left and bottom right) showing shift-invariance in x (bottom left) and z (bottom right) – here note the shape of the PSFs is the same for both points.

III. SHIFT-INVARIANT SYNTHESIS OF THE FOCAL STACK

A. Motivation

When imaged by a plenoptic camera, the volume can be estimated by generating a focal stack through computational refocusing. When objects are present within the volume, each focal plane within the stack will contain energy from out-of-plane features. This phenomenon can be described by a PSF that characterizes the energy spread or blur from each point in the volume. Existing refocusing algorithms are based on a continuous-space synthesis equation that is inherently shift-variant. The figures shown in the top row of Fig. 4 demonstrate how the PSF changes throughout a reconstructed volume when existing synthesis equations are used.

The underlying cause of this can be understood through careful examination of (2). This simple relationship offers a great deal of insight into the source of shift-variance in this system. Plotting this equation for a fixed $(\boldsymbol{x}, \boldsymbol{u})$ as a function of α is equivalent to tracing the ray from a given aperture sample $\boldsymbol{u}$ to microlens-plane coordinate $\boldsymbol{x}$ through image space. An example of this is shown in Fig. 5 (unscaled traces) for two rays associated with a single $\boldsymbol{u}$ coordinate and two different $\boldsymbol{x}$ coordinates. Note that both rays converge to the same point at the aperture. This is a direct result of the way the system is modeled. Recall that each pixel underneath a microlens represents a specific point of intersection on the aperture. Thus, the physical angle associated with that pixel varies depending on which microlens is being considered according to its displacement from the optical focal axis. That is, the angle of the rays in image space associated with a given aperture point vary as a function of location on the microlens plane, yielding an inherently shift-variant system. For the system to be shift-invariant, the rays associated with a given $\boldsymbol{u}$ sample must be parallel as shown in Fig. 5 (scaled traces), regardless of which microlens sampled it. That is, each $\boldsymbol{u}$ sample must represent a specific angular value independent of position on the microlens.

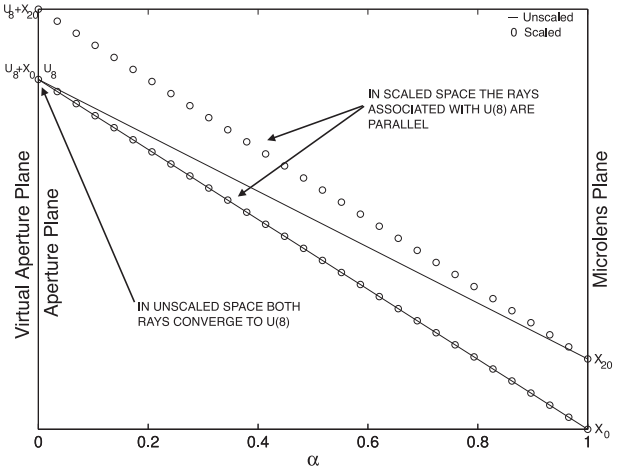


Fig. 5. Ray traces for rays associated with microlenses $\boldsymbol{x}_0$ and $\boldsymbol{x}_{20}$ and angular sample $\boldsymbol{u}_8$ in both unscaled and scaled coordinates. Note that in scaled space, both rays converge to the same point at the aperture, which indicates the angle associated with each $\boldsymbol{u}$ sample changes as a function of the $\boldsymbol{x}$ position at the microlense. Note that in scaled space both rays are now parallel, which indicates the angle associated with each $\boldsymbol{u}$ sample represents a fixed angle independent of the position $\boldsymbol{x}$.

A shift-variant form of the synthesis system is problematic because only shift-invariant systems can be modeled by convolution with a single kernel, and only convolution with a single kernel allows the system to be described by multiplication in the frequency domain. FFT-based deconvolution offers a fast and efficient means of removing volumetric blur but only from shift-invariant systems. Fortunately, under certain circumstances and with appropriate strategies, the overall continuous-space plenoptic acquisition-refocusing system can be closely approximated as shift-invariant. The figures shown in the bottom row of Fig. 4 demonstrate the desired shift-invariant form of the PSF.

In this section we propose a geometric transformation in the refocusing process that makes the continuous-space acquisition-refocusing system shift-invariant. In the process of this development we derive the form of the modified PSF that can be used in a Fourier-domain reconstruction algorithm.

B. Analysis of Existing Shift-Variant Refocusing

We demonstrate shift-variance of the standard refocusing algorithm by showing that the PSF—the resulting refocused image stack—varies with input point-source location. To begin, consider a point of light located in image space at $(\boldsymbol{x}_0, z_0)$ relative to the origin, where z_0 is the coordinate in the depth direction. This can be modeled by an impulse $A(z_0)\delta(\boldsymbol{x}_\alpha - \boldsymbol{x}_0, z - z_0)$, where $A(z_0)$ specifies the irradiance of the point source. The camera aperture limits the range of angular samples recorded; this is modeled by multiplying the impulse by a windowing function $W(\boldsymbol{u})$ that models the effects of the aperture. The light field for a point source at $(\boldsymbol{x}_0, z_0)$ in image space is defined by:

$$L_{F-z_0}(\boldsymbol{x}_{\alpha_0}, \boldsymbol{u}) = A(z_0)\delta(\boldsymbol{x}_{\alpha_0} - \boldsymbol{x}_0)W(\boldsymbol{u}) \tag{4}$$

where $\alpha_0 = \frac{F-z_0}{F}$.

To demonstrate shift-variance, consider the refocused image resulting from this light field formed from a point source at $(\boldsymbol{x_0}, z_0)$. First, we derive the light field at the microlens array by tracing along rays passing through $(\boldsymbol{x_0}, z_0)$. Let

$$\begin{aligned} L_F(\boldsymbol{x}, \boldsymbol{u}) &= L_{F-z_0}(\boldsymbol{x}_{\alpha_0}, \boldsymbol{u})\,|_{\boldsymbol{x}_{\alpha_0} = \alpha_0 \boldsymbol{x} - \boldsymbol{u}(\alpha_0 - 1)} \\ &= A(z_0)\delta\left(\alpha_0 \boldsymbol{x} - \boldsymbol{u}(\alpha_0 - 1) - \boldsymbol{x}_0\right) W(\boldsymbol{u}) \\ &= \frac{A(z_0)}{\alpha_0^2}\delta\left(\boldsymbol{x} - \frac{\boldsymbol{x}_0}{\alpha_0} - \boldsymbol{u}(1 - \frac{1}{\alpha_0})\right) W(\boldsymbol{u}) \end{aligned} \tag{5}$$

For the development in this section we set $A(z_0) = 1$ so that the impulse in image space has unit volume.

Forming a refocused image stack for all α is accomplished by substituting the light-field relationship (5) with $A(z_0) = 1$ into the irradiance equation (1)

$$\begin{aligned} E_{F-z}(\boldsymbol{x}_\alpha) &= \frac{1}{(F-z)^2}\int \frac{1}{\alpha_0^2}\delta\left(\boldsymbol{u}\left(1 - \frac{1}{\alpha}\right) + \frac{\boldsymbol{x}_\alpha}{\alpha} - \frac{\boldsymbol{x}_0}{\alpha_0}\right. \\ &\qquad \left. -\boldsymbol{u}\left(1 - \frac{1}{\alpha_0}\right)\right) W(\boldsymbol{u})\boldsymbol{du} \\ &= \frac{1}{\alpha_0^2\alpha^2 F^2}\int \delta\left(\frac{\alpha - \alpha_0}{\alpha\alpha_0}\boldsymbol{u} + \frac{\boldsymbol{x}_\alpha}{\alpha} - \frac{\boldsymbol{x}_0}{\alpha_0}\right) W(\boldsymbol{u})\boldsymbol{du} \\ &= \frac{1}{(\alpha - \alpha_0)^2 F^2} W\left(-\frac{\alpha_0}{\alpha - \alpha_0}\left[\boldsymbol{x}_\alpha - \frac{\alpha}{\alpha_0}\boldsymbol{x}_0\right]\right) \\ &= \frac{1}{(z - z_0)^2} W\left(\frac{F - z_0}{z - z_0}\left[\boldsymbol{x}_\alpha - \frac{F - z}{F - z_0}\boldsymbol{x}_0\right]\right) \end{aligned} \tag{6}$$

By inspecting this image stack function, we can see that when the input impulse shifts by $\boldsymbol{x}_0$, the image stack function shifts by a depth-dependent amount $\frac{F-z}{F-z_0}\boldsymbol{x}_0$, which skews the PSF. This is illustrated in Fig. 4, upper left plot, for two different input shifts in the x axis. Furthermore, the scale factor $F - z_0$ inside the aperture function $W()$ means that the PSF will spread out more when the point is at a smaller depth $F - z_0$ (Fig. 4, upper right). The result is that the refocused image stack is not shift-invariant either parallel to the focal plane or in depth.

C. Development of Shift-Invariant Image Synthesis Equations

Previous work [37] with light-field microscopy has acknowledged the shift-variant nature of the standard image synthesis equations and demonstrated that applying a single PSF for deconvolution nevertheless yields excellent results. However, a shift-invariant form of the synthesis equations remains desirable. This can be achieved through a different parameterization of the light-field to image-space mapping. Consider a new scaling relation between the coordinates on the microlens plane and the parallel coordinates on each refocused plane. We denote the modified refocused-plane coordinates as $\tilde{\boldsymbol{x}}_\alpha$. The new relationship is defined as:

$$\tilde{\boldsymbol{x}}_\alpha = \boldsymbol{x} - \boldsymbol{u}(\alpha - 1) \tag{7}$$

The relationship shown in (7) is subtly different from that of (2), but the impact is significant. In essence, this new mapping treats

the light field at the microlens plane as though it looks progressively larger as the refocused plane moves toward the aperture. An alternative viewpoint is that the new mapping adjusts the two-plane representation so that the $\boldsymbol{u}$ plane is now treated as a $\boldsymbol{u}+\boldsymbol{x}$ plane, as shown in Fig. 5. This view maintains the angles in the PSFs even as image points move away from the microlens plane or shift parallel to it. Fig. 5 illustrates the fact that the rays associated with a particular $\boldsymbol{u}$ sample no longer converge to the same point on the aperture in scaled-space. The scaling used to arrive at (7) forces the rays associated with a particular $\boldsymbol{u}$ to be parallel as shown in Fig. 5, which yields a shift-invariant system.

To demonstrate shift-invariance mathematically, we follow a similar procedure to the previous section. We begin with a shifted point source in image space and derive the resulting PSF. We show that this PSF varies only by a shift corresponding to the shift in the point source.

The light field in the modified coordinate system at an arbitrary focal plane z is denoted by $\tilde{L}_{F-z}(\tilde{\boldsymbol{x}}_\alpha, \boldsymbol{u})$. This can now be calculated in terms of the microlens-plane light-field pattern created by the previously defined impulse by using the fact that the light field is constant along rays in image space and using (7) to describe the relations in the modified coordinate system. Employing (5) with $A(z_0)=\alpha_0^2$ to normalize the impulse in scaled image space:

$$\begin{aligned}\tilde{L}_{F-z}(\tilde{\boldsymbol{x}}_\alpha, \boldsymbol{u}) &= L_F\left(\boldsymbol{x}, \boldsymbol{u}\right)\big|_{\boldsymbol{x}=\boldsymbol{u}(\alpha-1)+\tilde{\boldsymbol{x}}_\alpha} \\ &= \delta\left(\boldsymbol{u}\left(\alpha-1\right)+\tilde{\boldsymbol{x}}_\alpha-\boldsymbol{u}\left(1-\frac{1}{\alpha_0}\right)-\frac{\boldsymbol{x}_0}{\alpha_0}\right)W(\boldsymbol{u}) \\ &= \delta\left(\boldsymbol{u}\left[\alpha-1-1+\frac{1}{\alpha_0}\right]+\tilde{\boldsymbol{x}}_\alpha-\frac{\boldsymbol{x}_0}{\alpha_0}\right)W(\boldsymbol{u})\end{aligned} \tag{8}$$

Note that

$$\alpha-1-1+\frac{1}{\alpha_0}=-\frac{1}{F}\left(z-\frac{z_0}{\alpha_0}\right) \tag{9}$$

Defining the following transformed coordinates:

$$\tilde{\boldsymbol{x}}_{0,\alpha_0}=\frac{\boldsymbol{x}_0}{\alpha_0} \tag{10}$$

$$\tilde{z}_0=\frac{z_0}{\alpha_0} \tag{11}$$

we then substitute into the previous expression to obtain

$$\tilde{L}_{F-z}(\tilde{\boldsymbol{x}}_\alpha, \boldsymbol{u})=\delta\left(-\boldsymbol{u}\frac{1}{F}(z-\tilde{z}_0)+\tilde{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_{0,\alpha_0}\right)W(\boldsymbol{u}) \tag{12}$$

Note that evaluating this modified function at $z=\tilde{z}_0$ yields

$$\tilde{L}_{F-\tilde{z}_0}(\tilde{\boldsymbol{x}}_\alpha, \boldsymbol{u})=\delta(\tilde{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_{0,\alpha})W(\boldsymbol{u}) \tag{13}$$

which is the light-field function arising from an impulse $\delta(\tilde{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_{0,\,\alpha_0}, z-\tilde{z}_0)$ in scaled image space. The mapping to scaled image space requires that the scaling of the irradiance integral be modified to maintain equal energy across refocused image planes. The modified integral over the aperture is then calculated by integrating (12) over $\boldsymbol{u}$ to find modified refocused images at each z-plane:

$$\begin{aligned}\tilde{E}_{F-z}(\tilde{\boldsymbol{x}}_\alpha) &= \frac{1}{F^2}\int \tilde{L}_{F-z}(\tilde{\boldsymbol{x}}_\alpha, \boldsymbol{u})d\boldsymbol{u} \\ &= \frac{1}{F^2}\int \delta\left(-\boldsymbol{u}\frac{1}{F}(z-\tilde{z}_0)+\tilde{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_{0,\alpha_0}\right)W(\boldsymbol{u})d\boldsymbol{u}\end{aligned} \tag{14}$$

$$= \frac{1}{(z-\tilde{z}_0)^2}W\left((\tilde{\boldsymbol{x}}_\alpha-\tilde{\boldsymbol{x}}_{0,\alpha})\frac{F}{z-\tilde{z}_0}\right) \tag{15}$$

Thus, a shift of the input impulse by $(\tilde{\boldsymbol{x}}_{0,\alpha}, z_0)$ results in no change in the output focal stack other than a shift of the refocused image by the same amount, which demonstrates shift-invariance. This result also gives us a closed-form expression for the PSF of the overall system by setting $(\tilde{\boldsymbol{x}}_{0,\alpha}, \tilde{z}_0)$ to $(0,0)$. Equation (15) yields an expression for the overall system PSF:

$$h(\tilde{\boldsymbol{x}}_\alpha, z)=\frac{1}{z^2}W\left(\frac{F}{z}\tilde{\boldsymbol{x}}_\alpha\right) \tag{16}$$

The forward equation, neglecting noise and discretization, becomes

$$\tilde{E}_{F-z}(\tilde{\boldsymbol{x}}_\alpha)=h(\tilde{\boldsymbol{x}}_\alpha, z)*f(\tilde{\boldsymbol{x}}_\alpha, z) \tag{17}$$

where $f(\tilde{\boldsymbol{x}}_\alpha, z)$ is the light-source volume in the geometrically transformed coordinate system in image space.

The resulting algorithm for calculating a shift-invariant focal stack is

$$\tilde{E}_{F-z}(\tilde{\boldsymbol{x}}_\alpha)=\frac{1}{F^2}\int L_F\left(\boldsymbol{u}\left(\alpha-1\right)+\tilde{\boldsymbol{x}}_\alpha, \boldsymbol{u}\right)d\boldsymbol{u} \tag{18}$$

The impact of this modification to the algorithm can be seen by plotting the boundary of the PSF both before and after the scaling is applied. The boundary of the PSF at a given focal plane can be calculated by utilizing (2). The PSF can be plotted by setting u to the limits of the camera aperture, selecting two x intercept points at the microlens, and plotting $\tilde{x}_\alpha$ over a range of α values. The top row of Fig. 4 shows example PSFs depicting how the shape of the impulse response will change throughout the image in both x and z-directions. Equation (2) shows that the slope of the PSF boundary is set by $x-u$ which will vary for every point in the imaged space. The boundary of the PSF in scaled coordinates is obtained similarly from (7). The slope of the PSF boundary in scaled coordinates is now set by u, which is constant for every point throughout the volume. This can be seen in the second row of Fig. 4, which depicts the PSFs in scaled coordinates.

IV. Managing Other Sources of Shift-Variance

Unfortunately, sampling of the continuous-space system as well as the reconstruction process introduce further sources of shift-variance. These must be identified and removed or mitigated where possible to provide the best possible agreement between the shift-invariant continuous-space model and the shift-variant discretized system being analyzed. In the following sections, we identify sources of shift-variance due to sampling, and provide mitigating steps where possible. Though

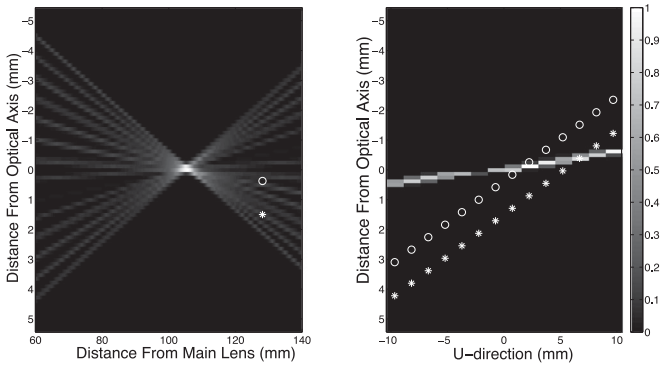


Fig. 6. PSF (left) showing banding with two sample points denoted by an $*$ and a $\circ$ in the PSF with the associated samples in the radiance array (right).

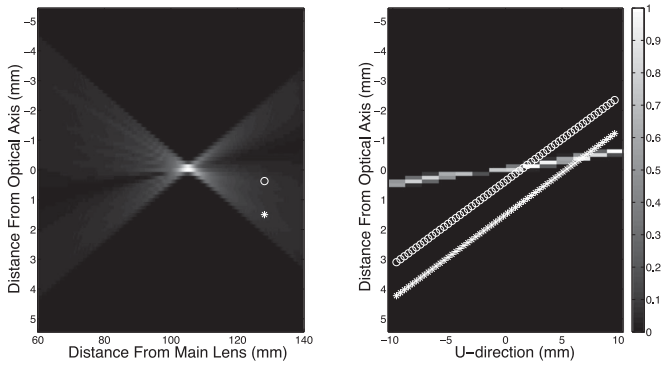


Fig. 7. PSF (left) obtained by oversampling the radiance array (right). Increasing the samples used in the reconstruction algorithm produces a smoother response.

some sources of shift-variance cannot be eliminated, experimental data is used to demonstrate the modified algorithm can be used to reduce blur in volumetric reconstructions.

A. *Effects of Discretization on Volume Reconstruction*

One source of spatial variability occurs when objects or features are imaged across multiple microlenses. This can occur when objects are located at the boundary of multiple microlenses or when objects off the focal plane are imaged. Discretization artifacts result in variations in the reconstructed points. Other sources of shift-variance result from the reconstruction algorithm itself and can be seen by inspecting the simulated PSF and its corresponding radiance array shown in Fig. 6.

The most prominent feature of the real-world PSF is a pattern of alternating light and dark bands that appear as the PSF decays from the optical centerline. The source of these bands becomes apparent when considering the samples of the radiance array responsible for example regions of the PSF. To illustrate, a point along one of the bright regions in the focal stack is denoted by an $*$. The samples integrated to obtain the point, likewise denoted by an $*$, are overlaid on the radiance array. Of note is the single sample that falls within the region of support (ROS) of sampled data in the radiance array. When the samples are summed, this sample is the most significant contributor to the intensity of the point in the PSF. A second point is chosen that falls within one of the dark bands and is denoted by a $\circ$ along with the corresponding samples from the radiance array. The samples for this point fall on either side of the ROS in the radiance array and therefore produce a much lower intensity in the corresponding point in the PSF.

The significance of this observation is that this banding is not the result of inadequate data collection, or aliasing, but of inadequately interpolating the data that has been recorded. To address the issue, at least in part, a simple increase in the density of samples along lines in the radiance array can significantly reduce the banding present in the focal stack. Fig. 7 depicts the case where four-fold interpolation is used to reconstruct the PSF from the same radiance array. Once again, the points in the PSF and the corresponding samples in the radiance array are identified.

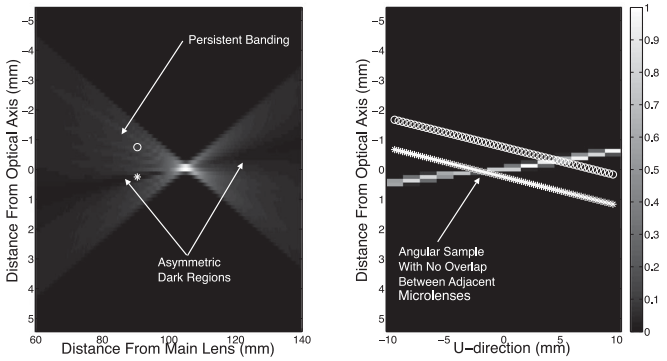


Fig. 8. Persistent asymmetric banding and dark regions in the PSF (left) with two example points identified and the corresponding samples from the radiance array (right) shown.

While interpolation improves the reconstruction significantly, some asymmetric banding and dark regions persist in the PSF. Both artifacts are identified in Fig. 8. Once again comparing the points in the PSF with the corresponding samples from the radiance array demonstrates the source of these artifacts.

First, the persistent banding is caused when the samples for a point in the PSF have the opposite slope compared to the radiance data. This can be seen when comparing the radiance array samples shown in Figs. 7 and 8. In the former, the discontinuities in the data ROS are masked by the overlapping angular information between adjacent microlenses. Because the slope of the samples has the same sign, more samples consistently fall on regions containing data, which results in a more uniform intensity. In the latter, the slope of the sampling path and the slope of the data ROS have opposite signs, and the data ROS has a stair-step structure. Thus, as the slope of the sampling path changes, the number of samples falling on the data ROS varies more dramatically. These variations cause the alternating light and dark regions in the refocused image. Because the banding is only significant when the slopes have opposite signs, banding appears only on one side of the optical focal plane.

Second, the dark regions on either side of the focused point are related to the persistent banding described above. The point denoted by an $*$ in Fig. 8 falls in one of the dark regions, and its corresponding samples are shown on the radiance array. It

can be seen that the samples cross a region in the data where no overlap exists between angular samples. As a result, fewer samples in the reconstruction fall within the data ROS, which results in a lower intensity.

These final artifacts are of particular interest because they are spatially variant. The persistent banding varies depending on the position of the point and is asymmetric about the point center, depending on which side of the optical focal plane the point appears. The dark regions appear when angular data is collected by a single microlens, which again is a spatially variant phenomena. Addressing these issues requires direct manipulation of the data or the imaging system and is the subject of future research. For the purposes of this work, these are sources of spatial variability within the reconstruction. However, the reconstructions included in this work performed on experimental data show that despite the remaining sources of shift-variance, deconvolution using the shift-invariant form of the continuous-space synthesis equation results in reduced volumetric blur.

V. Experiments

A. Basic Deconvolution Algorithm

With the imaging system modeled and sources of variability identified and addressed where possible, we approximate the forward model in the frequency domain as

$$Y(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z) = H(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z) X(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z) + U(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z) \tag{19}$$

where $Y(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z)$ is the Fourier transform of the refocused image $\tilde{E}_{F-z}(\tilde{\boldsymbol{x}}_\alpha)$ with respect to $(\tilde{\boldsymbol{x}}_\alpha, z)$, $X(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z)$ is the Fourier transform of the 3D light-source object, and $U(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z)$ represents observation noise and sources of modeling error. The PSF has sharp discontinuities at its boundaries due to the finite aperture. The result is a frequency response with zeros, which would unacceptably amplify the noise term if direct inversion is attempted. To limit noise amplification and PSF mismatch, we view the image as a random process and employ a simple form of the multidimensional Wiener filter given by

$$\hat{X}(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z) = \left[\frac{H^*(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z)}{|H^*(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z)|^2 + \frac{S_u(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z)}{S_x(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z)}} \right] Y(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z), \tag{20}$$

where S_u and S_x are the power spectrum of the noise and the original image volume respectively. When the power spectra are not known, as is the case here, the ratio of the power spectra can be approximated by a regularization parameter K, usually with little performance degradation. Substituting this into (20) gives

$$\hat{X}(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z) = \left[\frac{H^*(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z)}{|H^*(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z)|^2 + K} \right] Y(\tilde{\boldsymbol{\omega}}_{\boldsymbol{x}}, \omega_z), \tag{21}$$

Consider the computational complexity of this approach. For an image with $N \times N$ spatial samples, $M \times M$ angular samples, K focal planes and an oversampling factor J, the integral refocusing step requires $O(KJ^2M^2N^2))$ operations. Other approaches, such as FFT-based methods, may be developed that could drop this overhead considerably. The deconvolution algorithm itself requires $O(kN^2 \log_2(KN^2))$ operations for the 3D FFTs and $O(kN^2)$ operations for the regularized reconstruction step. Specific execution times are given in the discussion of the 3D experiment that follows.

B. Effects of Regularization and Artifact Identification

To demonstrate the efficacy of deconvolution when applied to shift-invariant plenoptic imaging and to identify artifacts in the reconstruction, two example cases are presented. The first is a simplified 2D case used to identify artifacts in the deconvolved reconstruction. The second is a 3D case used to demonstrate results when the algorithm is applied to experimentally acquired volumetric data.

C. 2D Reconstruction

For the 2D example case, the system is modeled with a native focal plane at 100 mm, and the remaining system parameters as identified in Table I. The simulations presented here utilize a microlens pitch that is not an integer multiple of the pixel pitch. As a result, some overlap exists between pixels associated with adjacent microlenses, which accurately simulates the physical scenario. These ambiguous pixels must be discarded before creating the radiance array.

Another important aspect of generating the radiance array is properly registering the microlens array with respect to the sensor. This step accounts for shifts and/or rotations of the microlens array relative to the sensor as well as a microlens pitch that is not an integer multiple of the pixel pitch. While the details of such a process are not considered here, the simulations provided here are calibrated via a calibration image in the same way experimental data would be calibrated. This process identifies the center of each microlens with respect to the sensor. The sensor data is then interpolated onto a uniform grid that simulates the perfectly aligned/integer-multiple scenario. This ensures that each row of the radiance array corresponds to the same angle while each column represents the same spatial location throughout the array. Without this step, each entry within the 2D array (for 2D images) or the 4D array (for 3D volumes) must be uniquely identified within the (x, u) space or the (x, y, u, v) space respectively. Utilizing a plaid—that is, separably sampled—array significantly improves computational efficiency of the refocusing algorithm. This plaid data is then used by the refocusing algorithm to generate the focal stack. Inclusion of these steps is intended to more accurately model the physical system, its parameters, and potential sources of error from the algorithm.

A PSF must be selected in order to apply deconvolution to a simulated point field. Section III-B and (7) provide a convenient means to quickly simulate a PSF based on the camera parameters. Equation (7) shows that the width of the PSF is set by the width of the camera aperture u and focuses to a point at the plane. The intensity falls off from the peak value based on the area of the disc at any given focal plane. Simulations and experimental results presented here are based on this ideal simulated PSF. The impacts of variations in the PSF and use of experimentally derived PSFs is the subject of future work.

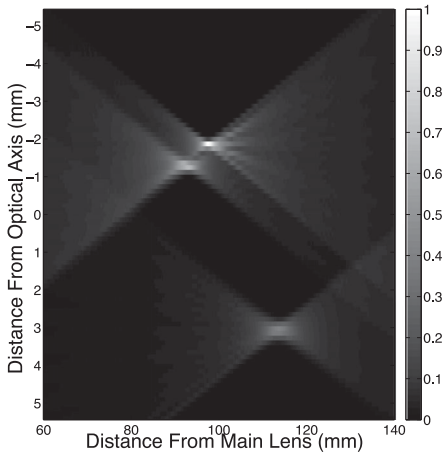


Fig. 9. Focal stack for three simulated points.

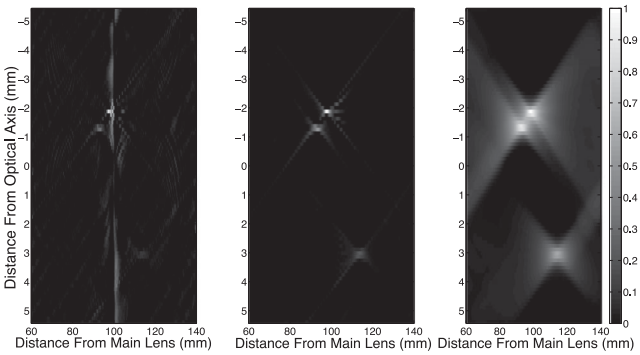


Fig. 10. Comparison of the deconvolution utilizing regularization values $K = 1 \times 10^{-7}$ (left), $K = 1 \times 10^{-4}$ (center), and $K = 0.1$ (right).

With a PSF chosen, the next step is to simulate an example field of point sources. A test case is selected with three points placed within the imaged space at $(1.3\,\text{mm}, 91.2\,\text{mm})$, $(1.9\,\text{mm}, 97.5\,\text{mm})$, and $(-2.5\,\text{mm}, 112.5\,\text{mm})$. The corresponding focal stack is then generated, and the image is reconstructed utilizing deconvolution. The resulting focal stack is shown in Fig. 9, which clearly depicts the blur resulting from out-of-plane points. Fig. 10 compares the results of the deconvolution when various levels of regularization are applied and highlights the importance of selecting the appropriate value for K. Too little regularization results in a poor reconstruction dominated by artifacts. Only two of the particles are easily identifiable as can be seen in the leftmost image of Fig. 10. Applying too much regularization increases blurring of the focal stack as shown by the rightmost image in Fig. 10. Selecting the appropriate level of regularization results in three clearly identifiable points with significantly reduced blur, as shown in the center image of Fig. 10. In practice, this parameter can be set experimentally and held constant as long as the imaging conditions remain the same from one acquisition to another.

Unfortunately, selection of the optimal regularization parameter does not ensure complete elimination of artifacts. These can be seen in the center image in Fig. 10. The deconvolved points display an X-like pattern because the limited angular view of the aperture makes the system noninvertible. Furthermore, the uppermost point displays the persistent banding in both the object response and the deconvolved image. These banding artifacts cannot be eliminated with deconvolution using a single PSF, as the banding phenomena is shift-variant. While techniques exist for addressing spatially variant PSFs [54], they increase the computational burden of the algorithm. Nevertheless, these results show a significant improvement over the estimate produced by simply generating a focal stack from the plenoptic data. As a result, we are able to achieve much improved depth resolution after deconvolution processing.

TABLE II
3-D SIMULATION PARAMETERS

Microlens Focal Length	f_l	0.500 mm
Microlens Pitch	p_l	0.125 mm
Pixel Pitch	p_p	0.0074 mm
Number of Pixels	n_p	1503×1503
Number of Microlenses	n_l	89×89
Main Lens Focal Length	f_m	50 mm
Volume Dimensions	(x, y, z)	$11.1\ \text{mm} \times 11.1\,\text{mm} \times 95.3\,\text{mm}$

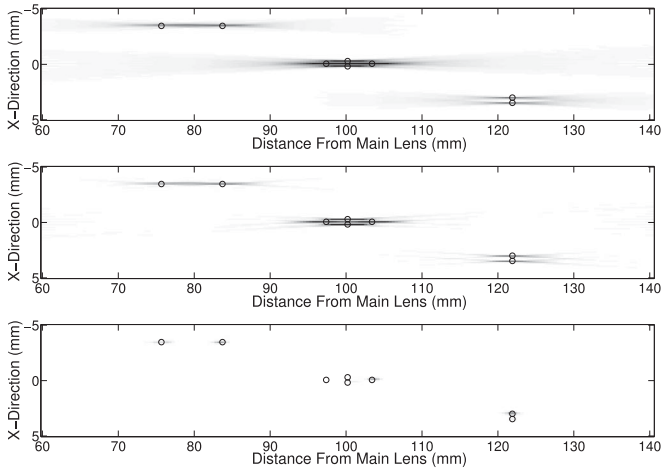


Fig. 11. Focal stack of simulated 3D volume (top), deconvolved focal stack (middle), and MART reconstruction of the simulated volume (bottom). Locations of simulated points indicated by O. Figures are presented in inverted grayscale (i.e., white $= 0$ and black $= 255$) and scaled individually for clarity.

D. Comparison of Deconvolution and MART

The 2D cases presented in previous sections have been used to demonstrate concepts and identify characteristics of the algorithm. In the following sections, the algorithm is extended to 3D volumetric scenes, and the reconstructions are performed via 3D deconvolution.

In order to provide a quantitative comparison of deconvolution to existing state-of-the-art reconstruction methods, the deconvolution algorithm presented here and an existing multiplicative algebraic reconstruction technique (MART) algorithm [40] were used to reconstruct a simulated 3D volume. Table II shows the simulation parameters, which were selected to mimic parameters of the physical imaging system used in subsequent testing. Because the depth at which images can be refocused clearly is limited by the angular resolution of the camera [9], [46], multiple spatial impulses were simulated. Fig. 11 (top)

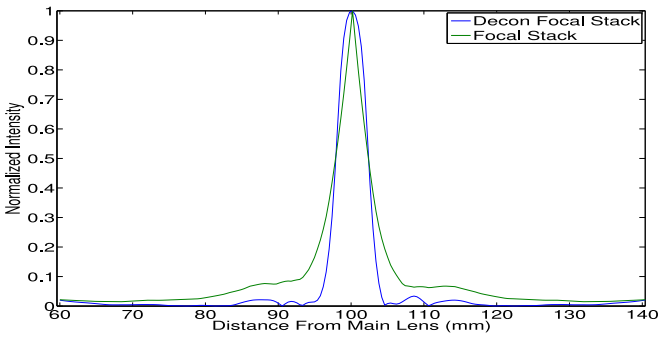


Fig. 12. The intensity of the focal stack and the deconvolved volume along the z-axis for the point centered at $(0.25\text{ mm}, 0\text{ mm}, 0\text{ mm})$ shown in Fig. 11.

shows the focal stack of the eight simulated points with the locations of the simulated points identified with the "O". This distribution was selected to test spatial resolution in both the $x-$ and z-directions for various locations throughout the volume. Performance in the y-direction will be the same as the x-direction for a system symmetric in the $x-$ and $y-$directions. Mean-squared error (MSE) is used to provide a quantitative metric for comparison.

Fig. 11 clearly demonstrates the differences between deconvolution and the MART reconstruction. The deconvolved volume reduces the image blur but does not eliminate it completely, though each spatial impulse is retained. In contrast, the MART algorithm reconstructed discrete points with residual blur restricted to the z-direction but failed to recover all points in the simulated volume. A comparison of mean squared error between the focal stack $(1.1e^{-3})$, the deconvolved volume $(6.9e^{-4})$, and the MART reconstruction $(1.5e^{-5})$ shows MART yielded the lowest error. However, there is a clear tradeoff between retaining each of the simulated points versus residual blur in the reconstruction. Because MART is multiplicative, it enforces a nonnegativity constraint, giving it a decided advantage when reconstructing isolated points. This advantage is somewhat limited in reconstructed distributed light sources, as in the next experiment.

Resolution is another metric commonly used for quantifying the improvements achieved by algorithms such as these. However, for the cases presented here, resolution is not only difficult to define, but also dependent on multiple parameters. To visualize resolution performance, Fig. 12 is obtained by plotting the intensity of the focal stack and the deconvolved volume along a line in the depth dimension through one of the points at the origin of Fig. 11. This figure shows that both the PSF and the deconvolved point decay rapidly away from the point-source origin, with the majority of the improvement due to deconvolution seen in the lower intensity levels. These large tails in the focal stack can become problematic in applications such as PIV, whereas these tails are suppressed considerably in the deconvolved image. Note that this is only a 1D slice of the 3D object. Returning to Fig. 11, it can be seen that the cross-range resolution at the origin of each point is nearly identical for both the focal-stack points as well as the deconvolved points. Moving away from the origin, the cross-range resolution of the deconvolved point improves in comparison to the focal stack. The rate of decay for the reconstructed point sources will also depend greatly on the regularization parameter used, which depends on the noise level. In addition, resolution will depend on the particular algorithm chosen. Exploration of these parameters in detail as well as their impact on reconstruction resolution are subjects of future work.

TABLE III
PLENOPTIC CAMERA PARAMETERS

Microlens Focal Length	f_l	0.500 mm
Microlens Pitch	p_l	0.125 mm
Pixel Pitch	p_p	0.0074 mm
Number of Pixels	n_p	3280×4904
Sensor Size		24 mm $\times$ 36 mm
Number of Microlenses	n_l	193×289
Main Lens Focal Length	f_m	100 mm

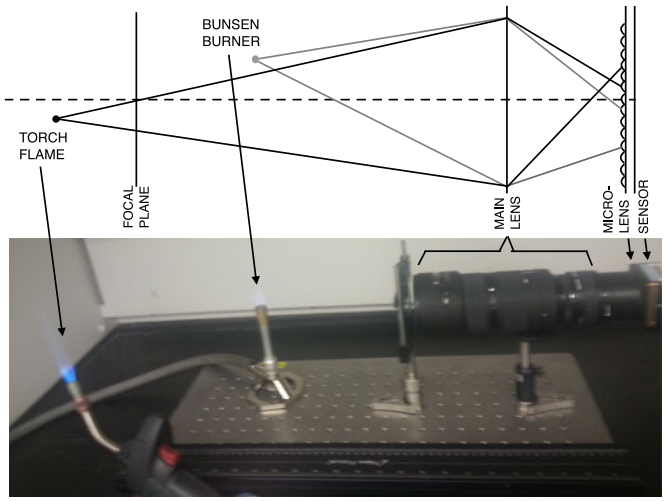


Fig. 13. Experimental setup showing the plenoptic camera and the two flames being imaged.

E. Volumetric Flame Reconstruction

Two experimental case are presented to demonstrate how the system characteristics and artifacts demonstrated in the previous sections impact volumetric reconstruction with a real-world system. The plenoptic camera parameters are presented in Table III. This configuration places the optical focal plane at 50 cm in front of the camera CCD, or 35.1 mm in front of the camera aperture. For the experiment, two flames are utilized as shown in Fig. 13. The first is from a Bunsen burner located to the right of centerline and 6 cm in front of the camera focal plane. The second is from a handheld torch located to the left of centerline and 10 cm behind the camera focal plane. The plenoptic camera is then used to capture an image of both flames burning. The sensor data is shown in Fig. 14. Both flames appear blurred in the raw image as neither flame is on the optical focal plane.

The deconvolution begins by creating a focal stack of both the flame image and the PSF. Fig. 15(a) shows an (x, y) slice of the volume, a single focal plane, at 44 cm where the Bunsen burner is in focus. This focal plane clearly shows the image blur resulting from the torch flame 16 cm further from the camera.

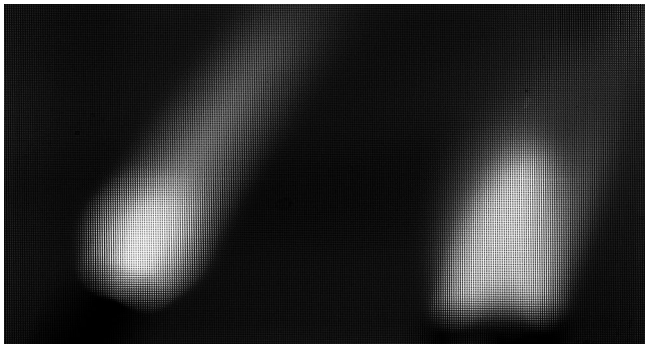


Fig. 14. Sensor data obtained by imaging the experimental setup—scaled for clarity.

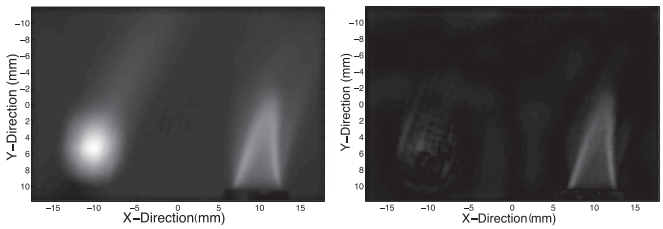


Fig. 15. Comparison slices of the (a) focal stack and the (b) deconvolved image along the (x, y) plane at $z = 44$ cm.

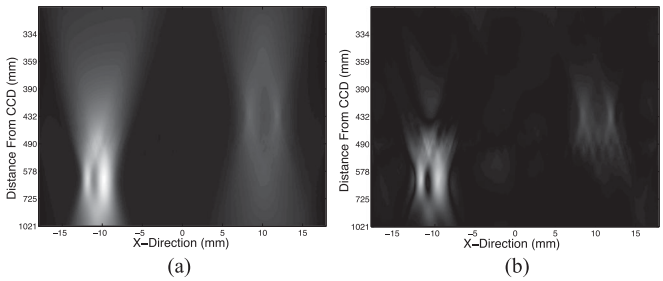


Fig. 16. Comparison slices of the (a) focal stack and the (b) deconvolved image along the (x, z) plane at $y = 6.1$ mm.

In a blur-free image, we would expect to see only the in-focus Bunsen burner flame on the right side of the image. An (x, z) slice through the flames shown in Fig. 16(a) provides another view demonstrating how the energy from each flame is spread into adjacent focal planes. Again, in the absence of image blur, each flame would be localized to the planes where the flame is present rather than smeared into adjacent planes as shown in the figure. Note that the scaling of the z-axis is the result of choosing a linear spacing between image-space samples, which results in a nonlinear sampling in object space. Sample spacing in object space is determined using the thin lens approximation

$$\frac{1}{f_m} = \frac{1}{s_i} + \frac{1}{s'} \tag{22}$$

where s_i is the image distance, s' is the object distance, and f_m is the main lens focal length.

Performing the deconvolution and then considering the same (x, y) and (x, z) slices as before shows a significant reduction in out-of-plane energy. Of particular note, the blur from the torch flame is almost completely eliminated in the (x, y) slice at 44 cm shown in Fig. 15(b), while the structure of the Bunsen burner flame is now well defined. Examining the same (x, z) slice as before shows how the image blur is drastically reduced in the z-direction in Fig. 16(b). Once again, the structure of the flames is well defined in-plane. The out-of-plane blur is not completely removed in the reconstruction, though it is significantly reduced.

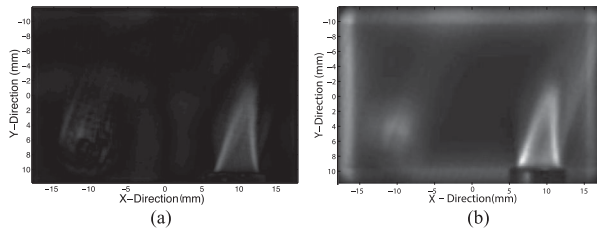


Fig. 17. Comparison slices of the (a) deconvolved image and the (b) MART reconstruction along the (x, y) plane at $z = 44$ cm.

In order to provide a comparison of these results to existing state-of-the-art methods, the MART algorithm was again used to reconstruct the volume. The (x, y) slice at 44 cm obtained from the MART reconstuction is shown in Fig. 17(b). Comparing this result to the deconvolution shown in Fig. 17(a) shows significantly more residual blur from the torch flame as well as reconstruction artifacts around the perimeter of the image. This shows that for a continuous field, such as the flame scenario used here, deconvolution may provide improved results when compared to MART.

The deconvolution for the 3D case presented required less than 10 seconds on a typical laptop without the benefit of parallel processing. In contrast, the MART algorithm used for comparison in this work required approximately 15 minutes to reconstruct the same volume. Unfortunately, the integral-based refocusing algorithm used to generate the focal stacks for both the example volume and the PSF is processor intensive and can require an order of magnitude more time to complete. This algorithm is easily parallelizable, and such algorithms have been demonstrated in our lab to generate similar focal stacks twice as fast as a non-parallel algorithm. Nonetheless, a faster method is desirable, and Fourier-based refocusing algorithms have the potential to significantly reduce the total time required to generate an estimate of the imaged volume. Recent results on this approach have been reported in [55]. The development of such algorithms is also the subject of the authors' future research.

A second example illustrates the value of volumetric reconstruction in seeing details that are ordinarily obscured by standard imaging. Fig. 18(a) shows an image of a test plate obscured by a flame in front of it. The image was obtained by calculating a focal stack and selecting the depth image that is focused on the test plate. Details behind the flame are not visible because of the flame intensity in front of the plate. A volume image was obtained using the proposed volume reconstruction algorithm. Then the image slice at the depth of the plate was extracted for comparison. This image is shown in Fig. 18(b). The images have been rescaled to obtain comparable scaling of the test plate. Much of the detail behind the flame has now been made visible.

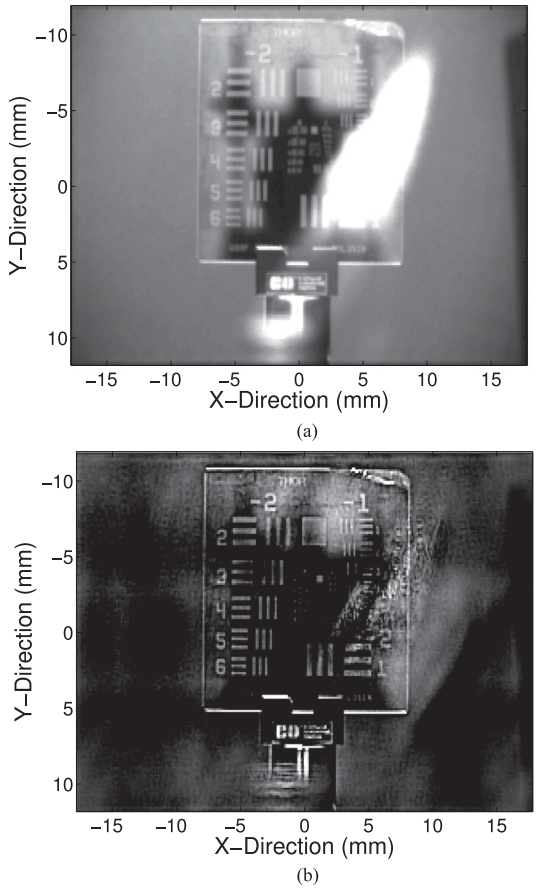


Fig. 18. (a) Comparison slices from an example focal stack and the (b) corresponding slice from the deconvolved volume. Details hidden by the flame in the focal stack are now visible in the deconvolved image.

VI. Conclusion

After developing a shift-invariant form of the refocusing algorithm, we have shown that deconvolution is an effective and computationally efficient means of volume reconstruction from plenoptic data. Given the PSF and focal stack, the volume can be reconstructed in seconds. When compared to existing tomographic methods such as MART which may require several hours to reconstruct modestly sized volumes [38]–[40], this is a significant improvement.

The quality of the reconstructions obtained via deconvolution is limited by the mismatch between the PSF and the object response as well as boundary artifacts. While the plenoptic imaging system impulse response can be modeled as shift-invariant, sampling and quantization effects result in a highly variable object response. As a result, a single PSF cannot provide a perfect reconstruction of the volume. Future work will explore filtering the acquired data and/or the reconstructed response and will develop iterative algorithms that can be used either exclusively or in conjunction with deconvolution to provide improved reconstructions. However, despite the residual blur, a fast convolution-based direct reconstruction is a significant improvement over the initial focal stack. We have demonstrated the value of the technique for a single example—reconstructing combustion events that could be used to improve burner design. This example illustrates the potential of plenoptic imaging for volume reconstruction in many other applications as well.

Acknowledgement

The authors would like to thank Tim Fahringer for his assistance in developing the integral-based algorithms utilized in this research and continued discussions regarding plenoptic imaging. They would also like to thank Kyle Johnson and Jeffrey Bolan for providing experimental data as well as for their work in developing the Light-Field Toolkit that aided in this research [56].

References

[1] J. Westerweel, G. E. Elsinga, and R. J. Adrian, "Particle image velocimetry for complex and turbulent flows," *Annu. Rev. Fluid Mech.*, vol. 45, no. 1, pp. 409–436, 2013. [Online]. Available: http://dx.doi.org/10.1146/annurev-fluid-120710-101204

[2] G. E. Elsinga, B. Wieneke, F. Scarano, and A. Schroder, "Tomographic 3D-PIV and applications," in *Particle Image Velocimetry*, vol. 112. Berlin, Germany: Springer-Verlag, 2008, pp. 103–125.

[3] H. C. Bheemul, G. Lu, and Y. Yan, "Three-dimensional visualization and quantitative characterization of gaseous flames," *Meas. Sci. Technol.*, vol. 13, no. 10, pp. 1643–1650, 2002. [Online]. Available: http://stacks.iop.org/0957-0233/13/i=10/a=318

[4] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Annu. Conf. Comput. Graph. Interactive Techn.*, ser. SIGGRAPH'96, 1996, pp. 31–42. [Online]. Available: http://doi.acm.org/10.1145/237170.237199

[5] J. Belden, T. T. Truscott, M. C. Axiak, and A. H. Techet, "Three-dimensional synthetic aperture particle image velocimetry," *Meas. Sci. Technol.*, vol. 21, no. 12, 2010, Art. no. 125403. [Online]. Available: http://stacks.iop.org/0957-0233/21/i=12/a=125403

[6] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. 23rd Annu. Conf. Comput. Graph. Interactive Techn.*, ser. SIGGRAPH'96, 1996, pp. 43–54. [Online]. Available: http://doi.acm.org/10.1145/237170.237200

[7] E. H. Adelson and J. Y. A. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 99–106, Feb. 1992. [Online]. Available: http://dx.doi.org/10.1109/34.121783

[8] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*. Cambridge, MA, USA: MIT Press, 1991, pp. 3–20.

[9] R. Ng, "Digital Light Field Photography," Ph.D. dissertation, Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Jul. 2006.

[10] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Stanford Univ., Stanford, CA, USA, Tech. Rep. CSTR 2005-02, Apr. 2005. [Online]. Available: http://graphics.stanford.edu/papers/lfcamera/

[11] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *Proc. 2009 IEEE Int. Conf. Comput. Photogr.*, Apr. 2009, pp. 1–8.

[12] F. Scarano, "Iterative image deformation methods in PIV," *Meas. Sci. Technol.*, vol. 13, no. 1, p. R1–R19, 2002. [Online]. Available: http://stacks.iop.org/0957-0233/13/i=1/a=201

[13] F. Scarano and M. L. Riethmuller, "Advances in iterative multigrid PIV image processing," *Exp. Fluids*, vol. 29, no. 1, pp. S051–S060, 2000. [Online]. Available: http://dx.doi.org/10.1007/s003480070007

[14] M. Raffel, C. E. Willert, S. T. Wereley, and J. Kompenhans, *Particle Image Velocimetry: A Practical Guide* (Experimental Fluid Mechanics), 2nd ed. New York, NY, USA: Springer-Verlag, Aug. 2007. [Online]. Available: http://www.springer.com/us/book/9783540723073

[15] R. J. Adrian and J. Westerweel, *Particle Image Velocimetry* (Cambridge Aerospace Series). Cambridge, U.K.: Cambridge Univ. Press, Dec. 2010. [Online]. Available: http://www.cambridge.org/us/academic/subjects/engineering/thermal-fluid s-engineering/particle-image-velocimetry?format=HB

[16] C. Brucker, D. Hess, and J. Kitzhofer, "Single-view volumetric PIV via high-resolution scanning, isotropic voxel restructuring and 3D least-squares matching (3D-LSM)," *Meas. Sci. Technol.*, vol. 24, no. 2, 2013, Art. no. 024001. [Online]. Available: http://stacks.iop.org/0957-0233/24/i=2/a=024001

[17] T. Hori and J. Sakakibara, "High-speed scanning stereoscopic PIV for 3D vorticity measurement in liquids," *Meas. Sci. Technol.*, vol. 15, no. 6, pp. 1067–1078, 2004. [Online]. Available: http://stacks.iop.org/0957-0233/15/i=6/a=005

[18] F. Pereira, M. Gharib, D. Dabiri, and D. Modarress, "Defocusing digital particle image velocimetry: A 3-component 3-dimensional DPIV measurement technique. Application to bubbly flows," *Exp. Fluids*, vol. 29, no. 1, pp. S078–S084, 2000. [Online]. Available: http://dx.doi.org/10.1007/s003480070010

[19] A. Schröder and C. E. Willert, Eds., *Particle Image Velocimetry: New Developments and Recent Applications* (Topics in Applied Physics). New York, NY, USA: Springer-Verlag, Mar. 2008. [Online]. Available: http://www.springer.com/us/book/9783540735274

[20] K. D. Hinsch, "Holographic particle image velocimetry," *Meas. Sci. Technol.*, vol. 13, no. 7, pp. R61–R72, 2002. [Online]. Available: http://stacks.iop.org/0957-0233/13/i=7/a=201

[21] T. Fahringer and B. Thurow, "Tomographic reconstruction of a 3-D flow field using a plenoptic camera," in *Proc. Fluid Dyn. Co-Located Conf.*, Jun. 2012. [Online]. Available: http://dx.doi.org/10.2514/6.2012-2826

[22] E. Deem, Y. Zhang, L. Cattafesta, T. Fahringer, and B. Thurow, "On the resolution of plenoptic PIV," *Meas. Sci. Tech.*, vol. 27, 2016, Art. no. 084003.

[23] K. Lynch and B. Thurow, "3-D flow visualization of axisymmetric jets at Reynolds number 6,700 and 10,200," *J. Vis.*, vol. 15, no. 4, pp. 309–319, 2012. [Online]. Available: http://dx.doi.org/10.1007/s12650-012-0141-2

[24] T. W. Fahringer, K. P. Lynch, and B. S. Thurow, "Volumetric particle image velocimetry with a single plenoptic camera," *Meas. Sci. Technol.*, vol. 26, no. 11, 2015, Art. no. 115201.

[25] B. Bohm, C. Heeger, R. Gordon, and A. Dreizler, "New perspectives on turbulent combustion: Multi-parameter high-speed planar laser diagnostics," *Flow, Turbulence Combustion*, vol. 86, no. 3/4, pp. 313–341, 2011.

[26] G. Kychakoff, K. Knapp, R. D. Howe, and R. K. Hanson, "Flow visualization in combustion gases using nitric oxide fluorescence," *AIAA J.*, vol. 22, no. 1, pp. 153–154, Jan. 1984. [Online]. Available: http://dx.doi.org/10.2514/3.8359

[27] K. Watson, K. Lyons, J. Donbar, and C. Carter, "Scalar and velocity field measurements in a lifted CH4-Air diffusion flame," *Combustion Flame*, vol. 117, no. 1/2, pp. 257–271, 1999.

[28] P. Kothnur, M. Tsurikov, N. Clemens, J. Donbar, and C. Carter, "Planar imaging of CH, OH, and velocity in turbulent non-premixed jet flames," *Proc. Combustion Inst.*, vol. 29, no. 2, pp. 1921–1927, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1540748902802334

[29] R. K. Hanson, "Applications of quantitative laser sensors to kinetics, propulsion and practical energy systems," *Proc. Combustion Inst.*, vol. 33, no. 1, pp. 1–40, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1540748910003913

[30] L. Ma *et al.*, "Tomographic imaging of temperature and chemical species based on hyperspectral absorption spectroscopy," *Opt. Express*, vol. 17, no. 10, pp. 8602–8613, May 2009. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-17-10-8602

[31] K. Y. Cho, A. Satija, T. L. Pourpoint, S. F. Son, and R. P. Lucht, "High-repetition-rate three-dimensional OH imaging using scanned planar laser-induced fluorescence system for multiphase combustion," *Appl. Opt.*, vol. 53, no. 3, pp. 316–326, Jan. 2014. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-53-3-316

[32] J. Hult *et al.*, "Quantitative three-dimensional imaging of soot volume fraction in turbulent non-premixed flames," *Exp. Fluids*,vol. 33, no. 2, pp. 265–269, 2002. [Online]. Available: http://dx.doi.org/10.1007/s00348-002-0410-2

[33] W. Fischer and H. Burkhardt, "Three-dimensional temperature measurement in flames by multispectral tomographic image analysis," *Proc. SPIE*, vol. 1349, pp. 96–105, 1990. [Online]. Available: http://dx.doi.org/10.1117/12.23521

[34] M. Hossain, G. Lu, and Y. Yan, "Three-dimensional reconstruction of combustion flames through optical fiber sensing and CCD imaging," in *Proc. 2011 IEEE Instrum. Meas. Technol. Conf.*, May 2011, pp. 1–5.

[35] T. Qiu, Y. Yan, and G. Lu, "A new edge detection algorithm for flame image processing," in *Proc. 2011 IEEE Instrum. Meas. Technol. Conf.*, May 2011, pp. 1–4.

[36] J.-B. Sibarita, *Deconvolution Microscopy*. Berlin, Germany: Springer-Verlag, 2005, pp. 201–243. [Online]. Available: http://dx.doi.org/10.1007/b102215

[37] M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz, "Light field microscopy," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 924–934, Jul. 2006. [Online]. Available: http://doi.acm.org/10.1145/1141911.1141976

[38] T. Fahringer, K. Lynch, and B. Thurow, "Volumetric particle image velocimetry with a single plenoptic camera," *Meas. Sci. Tech.*, vol. 26, 2015, Art. no. 115201.

[39] T. Fahringer and B. Thurow, "The effect of grid resolution on the accuracy of tomographic reconstruction using a plenoptic camera," in *Proc. 51st AIAA Aerosp. Sci. Meeting Including New Horizons Forum Aerosp. Expo.*, 2013. [Online]. Available: http://dx.doi.org/10.2514/6.2013-39

[40] T. Fahringer and B. Thurow, "Filtered refocusing: A volumetric reconstruction algorithm for plenoptic-PIV," *Meas. Sci. Tech.*, vol. 27, 2016, Art. no. 094005.

[41] M. Broxton *et al.*, "Wave optics theory and 3-d deconvolution for the light field microscope," *Opt. Express*, vol. 21, no. 21, pp. 25 418–25 439, Oct. 2013. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-21-21-25418

[42] Z. Mhabary, O. Levi, E. Small, and A. Stern, "Fast and exact method for computing a stack of images at various focuses from a four-dimensional light field," *J. Electron. Imag.*, vol. 25, no. 4, 2016, Art. no. 043002.

[43] J. P. Anglin, S. J. Reeves, and B. S. Thurow, "Efficient volumetric estimation from plenoptic data," in *Proc. SPIE Electron. Imag. Symp., Comput. Imag. XII*, San Francisco, CA, USA, Feb. 2014, Paper 90200T.

[44] M. Levoy, "Light fields and computational imaging," *Computer*, vol. 39, no. 8, pp. 46–55, Aug 2006.

[45] M. Levoy, "The digital Michelangelo project," in *Proc. 2nd Int. Conf. 3-D Digit. Imag. Model.*, 1999, pp. 2–11.

[46] R. Ng, "Fourier slice photography," in *Proc. ACM SIGGRAPH 2005 Papers*, 2005, pp. 735–744. [Online]. Available: http://doi.acm.org/10.1145/1186822.1073256

[47] B. Wilburn *et al.*, "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, Jul. 2005. [Online]. Available: http://doi.acm.org/10.1145/1073204.1073259

[48] V. Vaish *et al.*, "Synthetic aperture focusing using a shear-warp factorization of the viewing transform," in *Proc. 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2005, pp. 129–134. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.537

[49] I. Ihrke, J. Restrepo, and L. Mignard-Debise, "Principles of light field imaging: Briefly revisiting 25 years of research," *IEEE Signal Process. Mag.*, vol. 33, no. 5, pp. 59–69, Sep. 2016.

[50] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.

[51] T. G. Georgiev and A. Lumsdaine, "Resolution in plenoptic cameras," in *Proc. Front. Opt. 2009/Laser Sci. XXV/Fall 2009 OSA Opt. Photon. Tech. Dig.*, 2009, Paper CTuB3. [Online]. Available: http://www.opticsinfobase.org/abstract.cfm?URI=COSI-2009-CTuB3

[52] T. Georgeiv, K. C. Zheng, B. Curless, D. Salesin, S. Nayar, and C. Intwala, "Spatio-angular resolution tradeoff in integral photography," in *Proc. Eurograph. Symp. Rendering*, 2006, pp. 263–272.

[53] K. Lynch, "Development of a 3-D fluid velocimetry technique based on light field imaging," Master's thesis, Dept. Aerospace Eng., Auburn University, Auburn, AL, USA, May 2011.

[54] M. Temerinac-Ott, O. Ronneberger, R. Nitschke, W. Driever, and H. Burkhardt, "Spatially-variant Lucy-Richardson deconvolution for multiview fusion of microscopical 3D images," in *Proc. 2011 IEEE Int. Symp. Biomed. Imag., Nano Macro*, 2011, pp. 899–904.

[55] W. Fu, X. Tong, C. Shan, S. Zhu, and B. Chen, "Implementing light field image refocusing algorithm," in *Proc. 2nd Int. Conf. Opto-Electron. Appl. Opt.*, Oct. 2015, pp. 1–8.

[56] J. Bolan, E. Hall, C. Clifford, and B. Thurow, "Light-field imaging toolkit," *SoftwareX*, vol. 5, pp. 101–106, 2016.



Paul Anglin received the M.Sc. degree in electrical engineering from the University of Kansas, Lawrence, KS, USA, where he studied radar transmit coding and multi-element antenna array design. He also received the Ph.D. degree in electrical engineering from Auburn University, Auburn, AL, USA, where his interests included plenoptic imaging and digital image processing. He is currently a Principal Design Engineer at Tetrad Engineering, LLC, Valley, AL, USA.



Stan Reeves received the Ph.D. degree from the Georgia Institute of Technology, Atlanta, GA, USA, in 1990. He is currently a Professor in the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA. He was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the IEEE SIGNAL PROCESSING LETTERS. His research interests include MRI reconstruction and acquisition, image restoration and reconstruction, and optimal color and multispectral image acquisition.



Brian Thurow received the B.S, M.S., and Ph.D. degrees in mechanical engineering from The Ohio State University, Columbus, OH, USA, in 1999, 2001, and 2005, respectively. He is currently the Department Chair and a W. Allen and Marta Reed Professor in the Department of Aerospace Engineering, Auburn University, Auburn, AL, USA. He founded and directs the advanced flow diagnostics laboratory, which focuses on the development and application of novel optical diagnostic techniques for fluid dynamic measurements. His past work includes the development of MHz rate pulse burst laser technology for ultrahigh speed and scanning 3-D flow visualization. His research interest focuses on the development of light field imaging hardware and software for 3-D flow field imaging using simplified experimental arrangements. Applications include 3-D imaging of turbulent boundary layers, shock waves, unsteady separating flows, sprays, flames, and biomedical flows.

A Two-Phase 3-D Reconstruction Approach for Light Microscopy Axial-View Imaging

Yuanhao Guo, Yunpeng Zhang, and Fons J. Verbeek

***Abstract*—Three-dimensional representations in light microscopy are important for accurate shape assessment of model systems in biosciences. The computational multiview 3-D reconstruction seems feasible in obtaining the 3-D representations in particular for high-throughput. The specimen for imaging can have properties, i.e., transparency and translucency, that impede the detection of well-defined boundaries. Consequently, 3-D reconstruction and measurements, i.e., volume and surface area will be inaccurate. The motivation in this paper is therefore to develop a two-phase 3-D reconstruction approach for light microscopy axis-view imaging that can deal with these properties. In phase I of this approach, we develop an improved 3-D volumetric representation defined as the confidence map. This is derived from texture-augmented axial-view images of the specimen. In phase II, the 3-D reconstruction is accomplished by searching the optimal surface for the specimen over the confidence map. Subsequently, from the obtained 3-D reconstruction, 3-D measurements can be extracted. We present a high-throughput axial-view imaging architecture in light microscopy based on the vertebrate automated screening technology device. Using this imaging architecture, we present three typical datasets using different imaging modalities, which includes zebrafish larvae in bright-field and zebrafish liver in fluorescence. In the experiments, we have applied our approach on these datasets. We find that our approach yields precise 3-D shape representation and natural visualization. In comparison with a groundtruth setup, we have obtained accurate 3-D measurements both for the organism and the organ, which holds a promising shape assessment for model systems in biosciences.**

***Index Terms*—Axial-view imaging, confidence map and level set, light microscopy, 3D measurements, two-phase 3D reconstruction.**

I. Introduction

In MODERN life-science research, large volumes of microscopy data, i.e., representative images of a specimen from cellular to whole-mount scale are used. Robust and reproducible methods for data acquisition, image processing, and analysis are essential for further handling of the data. As data volumes get larger, comprehensive visualizations summarizing the data content also become important.

Images from samples, i.e specimens, acquired from light microscopy carry a lot of information that can be expressed by features such as shape and texture. Variations in these features provide information for a classification according to a specific condition. Such approach is typically employed in imaging applications in the fields of cytology, toxicology, oncology and others. At a glance, experts are able to observe variations in color; though subtle differences of size and shape are more difficult to capture. Moreover, as the volume of the data increases, it is not possible to objectively classify the differences. Therefore, these features need to be extracted in an automated manner and more advanced systems need to be developed [1]. For high-throughput systems this is especially true. Manual inspection is not feasible and in order to make inferences from the data, robust methods are required that produce accurate and reproducible measurements for shape analysis.

For the study of whole specimens, we should acknowledge that the nature of the shape is, de facto, in 3D. Thus, from a 3D shape representation of the specimen one will be able to produce an informative shape description, e.g., volume, surface area and 3D shape factors. With such features, we can accurately assess and compare shape variations in applications. These descriptors are also important for phenotypical systems evaluation that requires volume normalization into metric measurements [2].

3D images can be generated by different techniques. With a confocal laser scanning microscope (CLSM) 3D images can be acquired from which 3D representations can be derived. For high-throughput imaging, CLSM is less efficient and also larger objects are not feasible for CLSM imaging. The post-processing of the images is not straightforward [3]. Alternative to CLSM is the optical projection tomography (OPT) [4]. However, a serious drawback of this technique is that it does not work for live specimens and the specimen preparation is rather time consuming. In light field microscopy [5], microlenses are configured between the main lens and image sensor. Through 3D deconvolution a larger focal range is addressed and in this manner focal images are combined into a 3D image [6]. This technique is suitable for semi-transparent objects. In our research, we deal with specimens that consist of opaque, specular and transparent regions. Another limitation of light-field microscopy is the trade-off between spatial and angular resolution.

In this work, we present an axial-view imaging architecture based on light microscopy using the Vertebrate Automated Screening Technology (VAST-BioImager) [7]. From this

Manuscript received January 12, 2017; revised April 30, 2017; accepted July 7, 2017. Date of publication July 24, 2017; date of current version October 23, 2017. This work was supported by the Netherlands Organisation for Scientific Research (NWO) under Grant 834.14.001. The work of Y. Guo was supported by the China Scholarship Council (CSC) to participate in the Ph.D. programme of Leiden University. The guest editor coordinating the review of this paper and approving it for publication was Dr. Lu Fang. *(Corresponding author: Yuanhao Guo.)*

The authors are with the Imaging and Bioinformatics Group, Leiden Institute of Advanced Computer Science, Leiden University, 2333CA Leiden, The Netherlands (e-mail: y.guo.3@liacs.leidenuniv.nl; tjzhangyunpeng@gmail.com; f.j.verbeek@liacs.leidenuniv.nl).

Digital Object Identifier 10.1109/JSTSP.2017.2731742

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

imaging architecture, we can generate a series of 2D axial-view images for the specimen. In fact, this imaging belongs to the class of multi-view imaging techniques, which is also referred to as turn-table sequence [8]. In the field of computer vision, multi-view stereo (MVS) approaches have been developed to recover a 3D scene, or an object, from a range of 2D multi-view images. In these approaches, surface points, or a depth map, of an object can be estimated through the matching of correlated images [9]. For light microscopy imaging, however, objects of interest can be (partially) translucent and/or transparent; this holds for the microscope modalities that we are using, i.e., bright-field and fluorescence. These qualities are in favour of the observation of internal structures of the specimen but prevent the feasibility of surface points matching which is required for a MVS approach.

In our previous work, we have obtained a binary 3D volumetric representation from the shape-based 3D reconstruction which is intuitive and suitable for the goal of shape analysis through 3D measurements [10]. This method requires precise 2D shape segmentation. However, the accurate shape segmentation is sometimes difficult to obtain from images depicting specimens with poor-defined boundaries as a consequence of translucent and transparent properties of the specimen. Instead, a probabilistic framework for the 3D volumetric representation generates more flexible but not always accurate 3D shape description. To address these problems, we propose a **two**-phase approach for the **3D** reconstruction and measurements from **light** microscopy **axial**-view imaging; this is abbreviated as 2-3DLA. A schematic representation of this approach is shown in Fig. 1. The system takes a series of axial-view images as input and reconstructs a precise 3D model of the object from which accurate 3D measurements, i.e., volume and surface area, can be derived.

Specifically, in **Phase I** of 2-3DLA, an improved 3D volumetric representation in the form of a confidence map is constructed. First, the mean shift algorithm [11] is applied to improve the texture representation of the original images. Thus, the translucent and transparent regions of the specimen are enhanced so that they become more separable. Subsequently, approximations of the 2D shapes of the specimen are obtained and used for camera system calibration. Together with all the data, based on the camera projection intersection model, we obtain the confidence map by imposing a score, instead of a binary value or a probability, to each voxel element in the 3D space. The score is obtained by integrating two probabilistic models. These probabilistic models are jointly estimated from all the axial-view images in a multi-scale fashion and aims to generalize the texture distributions of the object and background. The confidence map indicates the likelihood of each voxel in 3D space to be part of the object. This is a more flexible 3D representation for the optimization of the surface. In **phase II**, under the assumption that the optimal surface which includes the specimen will be able to maximally separate the voxels from the object to the background as well as retain a smooth appearance of the object. The 3D reconstruction and measurements are accomplished by searching for such an enclosed surface over the confidence map. Consequently, we formulate the 3D reconstruction as a 3D segmentation problem which can be solved by employing a region based level set method.

In modern life-science, i.e., developmental biology and pharmacokinetics, zebrafish are widely used as model systems in various experimental settings [12]. Zebrafish are small in embryonic stages and can be easily studied with different types of microscopes; zebrafish are transparent in the early stages. In recent years, zebrafish are augmented with a large amount of reporter lines and these lines are extensively used in disease studies. We use our light microscopy axial-view imaging architecture to acquire three representative datasets including bright-field images of zebrafish and fluorescence images of zebrafish liver. We use the zebrafish and the corresponding axial-view images as a case study for the evaluation of 2-3DLA. We have found that the 2-3DLA can be successfully applied in this research field; we have achieved promising results for 3D reconstruction and measurements of zebrafish and its liver.

We summarize our major contributions as follows.

I We present a multi-modal axial-view imaging architecture using light microscopy.

II We present three representative datasets of light microscopy axial-view imaging including the zebrafish in bright-field and the zebrafish liver in fluorescence.

III We propose a computational and automated system named 2-3DLA to solve the problem of accurate 3D reconstruction and measurements.

IV The proposed 2-3DLA is applicable in the datasets to obtain detailed 3D shape description both for the zebrafish and its liver. This makes the 2-3DLA generic for shape analysis on the level of the organism as well as on the level of the organ.

The remainder of the paper is structured as follows. In Section II, the background and related topics of our approach is introduced and the proposed system is motivated. In Section III, we present the architecture of our high-throughput axial-view imaging and introduce three different complementary datasets. In Section IV, we describe the prior knowledge of the light microscope camera calibration and the binary 3D volumetric representation. This is followed by a detailed description of the proposed system. In Section V we present experiments and discuss the results. Finally, in Section VI we summarize, present conclusions and future work.

II. Background and Related Work

Given a sequence of calibrated multi-view images of an unknown scene, one category of the multi-view stereo (MVS) approaches aims to estimate a depth map for each view by matching each pixel of a binocular image pair [13]–[15]. The matching criterion is usually defined within a support window, i.e., a local neighbourhood, centered around a target pixel [16], [17]. Another approach formulates the depth map estimation as a continuous optimization problem of an energy functional integrating color, spatial consistency of neighbouring views and a global smoothness constraint [18]. A merging strategy is

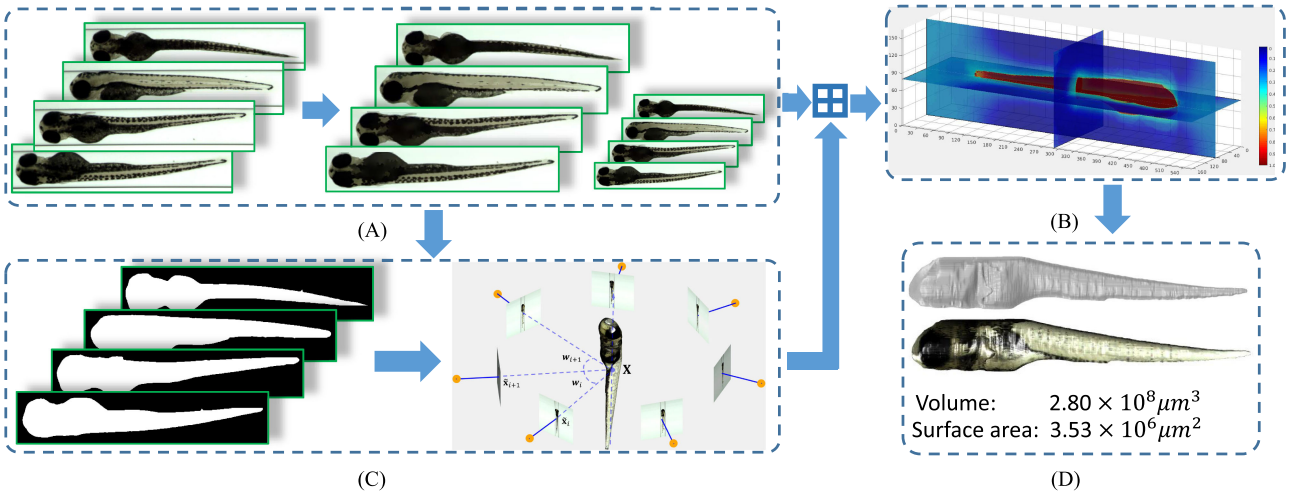


Fig. 1. A schematic of the 2-3DLA approach. Phase I: (A) A series of axial-view images of a specimen are acquired. The textures are augmented by the mean shift filtering and the multi-scale images are used as input of the system. (B) 2D shape approximations of the specimen are obtained from the augmented textures, with which the camera calibration can be accomplished. (C) An improved 3D volumetric representation in the form of a confidence map is derived from the textures through the volume intersection model. Phase II: (D) Region based level set method is adapted and applied to the confidence map to estimate the optimal enclosed surface retaining a natural shape and smooth appearance of the specimen, from which the volume and surface area are derived.

developed to integrate the multi-view depth maps. A patchmatch method [19] takes the slanted planar surface into account for a better depth map estimation in binocular stereo. An extension of the patchmatch method is developed for massive parallelization and integration of multi-view depth maps [20].

Alternatively, a conventional MVS approach directly estimates the surface points of an object using a sparse-to-dense strategy [21]. A novel approach extends the patch-based method into multi-view 4D reconstruction solving the problem of temporally consistent 3D modelling [22].

Regarding MVS, an accurate estimation of epipolar geometry is essential, which could significantly reduce the search space on the epipolar line and thereby increase the matching quality [23]. Structure from motion (SFM) enables the estimation of camera poses from the cooperation of salient point detection [24], [25] and bundle adjustment [26], [27]. However, both SFM and MVS rely on the quantity and quality of the salient point detection and matching from correlated images. In our light-microscopy imaging this is a challenge as object surface (and boundaries) are not always well-defined. In most cases, volume instead of surface for an object is observable.

For image based 3D reconstruction, the volumetric representation approach is an important category. The goal is to estimate a convex hull in the 3D space represented as discrete voxels according to their visibility to each view [8], [28]–[30]. Accordingly, the shape-based method defines the 3D object through the intersection of a set of projections exposed from the 2D shapes of the objects in the images. A space carving algorithm aims to recover the 3D object by wiping out the voxels which are consistently invisible to the views [31], [32]. These methods, however, require accurate image segmentations which are not always available in light microscopy imaging.

In addition, the textures from the multi-view images can be used to optimize the 3D volumetric representation based 3D reconstruction. [33]–[35]. Instead of using a binary representation, the probabilities indicating the membership of each voxel are estimated by applying multivariate Gaussian kernels on the textures of foreground and background, respectively. However, this requires user specification. Consequently, the quality of user input will, to a certain extent, determine the quality of the 3D reconstruction. Other methods attempt to fuse the shape and texture consistency in an integral deformable framework [36].

There are also other well-designed 3D reconstruction methods. Some in particular solve the problem of specular and transparent object reconstruction [37]. These methods work well on a macroscopic scale as they position special patterns behind the object to enable the shape-from-distortion or set up various lighting to collect surface reflective-highlights [38]–[40]. A recent study [41] presents a semantic reconstruction as a convex-relaxation formulation which combines a data term and a regularization constraint, achieving elaborate results on public datasets [14]. Nowadays, deep learning is used in the MVS to improve the matching quality [42], [43]. This type of method requires a large volume of training data, which, in our case, are not available.

From the study of related work, we conclude that the volumetric representation based approaches such as shape-based 3D reconstruction with binary volumetric representation is more promising in addressing the challenge of 3D reconstruction and measurements for light microscopy imaging. However, it is not sufficient, as in some cases accurate 2D shapes cannot be obtained. Therefore, we have developed the 2-3DLA method to solve the problem. In our approach, we first estimate a confidence map using the augmented textures from the axial-view images. The confidence map is actually an improved 3D representation. Subsequently, we accomplish the 3D reconstruction using the region based level set method on the confidence map.

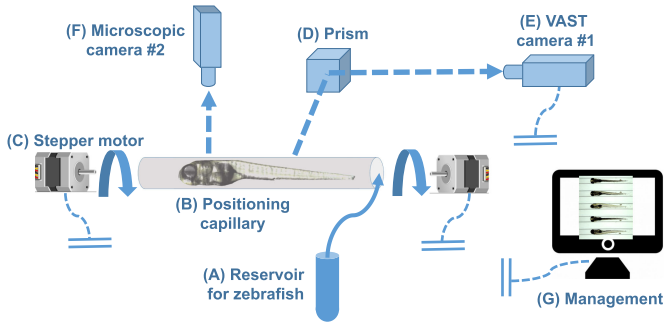


Fig. 2. A schematic drawing of light microscopy axial-view imaging based on the VAST-BioImager. A pump system loads one specimen at a time from the reservoir (A) and delivers it into the capillary (B) that keeps it in position. A set of stepper motors (C) drive the capillary motion in a full revolution. Through a prism (D), the VAST-camera (E) monitors and keeps the object in its field of view. Once the specimen is detected, the microscope camera (F) is able to acquire images of the specimen in any axial-view. The VAST-camera (E) manipulated by the management software (G) can also acquire images of the specimen while revolving presents.

A validation has shown that our method yields accurate 3D measurements.

III. Light Microscopy Axial-View Imaging and Dataset Collection

The VAST-BioImager is specifically set up to work for high-throughput imaging of zebrafish. For the remainder of the paper, the concept of specimen and zebrafish are both used. Specimen is used in a generic context while zebrafish is used in the particular application of the zebrafish high-throughput imaging. In the VAST-BioImager, the specimens, i.e., zebrafish larvae, are positioned along their longitudinal axis as in this manner the most important features can be readily observed; it is also related to the manipulation of the specimen in the capillary. In this orientation we obtain images from all axial-views along the profile axis from one full revolution. A schema of the imaging architecture is depicted in Fig. 2. A positioning module consists of a capillary (the holder of specimen) and a set of stepper motors that accomplish the rotation manipulation of the specimen.

A VAST-camera ($\#1$) is mounted with the device and used to detect the location and orientation of the specimen so as to keep it in the field of view for image acquisition. An object is revolved over 360 degrees by the stepper motors so that bright-field images for the specimen in the axial-views can be acquired. These image always depict the whole specimen. The VAST-camera ($\#1$) is an Allied Vision Systems, Pro Silica GE 1050 CCD (pixel size 5.5 μm $\times$ 5.5 μm.). This camera acquires images of 1024 $\times$ 1024 pixels. The resolving power for this camera system is about 13.4 μm.

The VAST-unit is mounted on a microscope, of which the microscope-camera ($\#2$) is able to acquire detailed microscopic images both in bright-field and fluorescence from arbitrary axial positions; this setup uses the Leica DFC450C CCD (pixel size 3.4 μm $\times$ 3.4 μm) and it acquires images of 1920 $\times$ 2560 pixels (5 Megapixel). We use two objectives in this case. The resolving power for a 2.5$\times$ objective is about 4.8 μm and for a 4$\times$ lens it is around 3.1 μm.

The image acquisition protocol is as follows. In a full revolution of the capillary, 84 axial-view images are acquired for the specimen. The step size between adjacent axial-views is around 4.3° ($360° \div 84 \approx 4.3°$). We have shown that 21 evenly sampled axial-views ($N = 21$) are sufficient to obtain accurate results [10]. We apply this acquisition protocol to obtain 3 different datasets that we will use as case-studies for the evaluation of our 2-3DLA approach under different imaging conditions.

Dataset A is obtained by the VAST-camera. A number of zebrafish larvae during different development stages, i.e., 3, 4 and 5 days post fertilization (dpf) are acquired. The dataset finally contains 12 examples for 3 dpf, 24 examples for 4 dpf and 24 examples for 5 dpf, respectively, and 60 examples in total. We will use the images from this dataset as example to interpret our 2-3DLA system. Examples of the images in this dataset are shown in Fig. 4.

Dataset B is obtained using the microscope-camera; the VAST-unit manipulates the position of the specimen while, in this case, the microscope-camera acquires bright-field images. For the acquisition the objective 2.5 $\times$ /0.07 NA is used. *Dataset B* consists of a set of representative examples, i.e., 3 specimens of 5 dpf. Examples of the images are depicted in Fig. 7.

Dataset C represents a collection of images of an internal structure of the zebrafish, i.e., an organ. These images are obtained using the fluorescence imaging modality as the organ is specifically visible with fluorescence. For fluorescence we use the objective 4 $\times$ /0.12NA to collect the dataset for the zebrafish liver. We used 7 zebrafish samples of 3 dpf. The microscope is equipped for fluorescence and the images are acquired with the microscope-camera. Examples of the images are depicted in Fig. 8; in Appendix A the preparation of these samples is given.

IV. Two-Phase 3D Reconstruction From Axial-Views

In this section we elaborate on the 2-3DLA approach. For the 3D reconstruction using axial-views, a microscope camera calibration is necessary. This can be solved by the algorithm of voxel residual volume maximization. As a baseline method, the shape-based 3D reconstruction is introduced and used for comparison with the our method.

A. Microscope Camera System Calibration

For a successful operation of the multi-view 3D reconstruction, an accurate camera system calibration is essential. Standard algorithms [44] tackle the problem by imaging regular-shaped patterns from multiple poses and solving a linear regression or estimating the epipolar geometry by matching salient points from correlated images. For an imaging system operating on micrometer-scale, this is, however, not feasible. Inspired by the idea of silhouette coherence maximization [45], we propose to estimate the pinhole imaging geometry by the algorithm of voxel residual volume maximization (VRV) [10] which can be efficiently implemented. To this end, we need 2D shape approximations which can be obtained by the mean shift (MS) algorithm. The MS algorithm improves the texture representation of the object so that the translucent and transparent regions

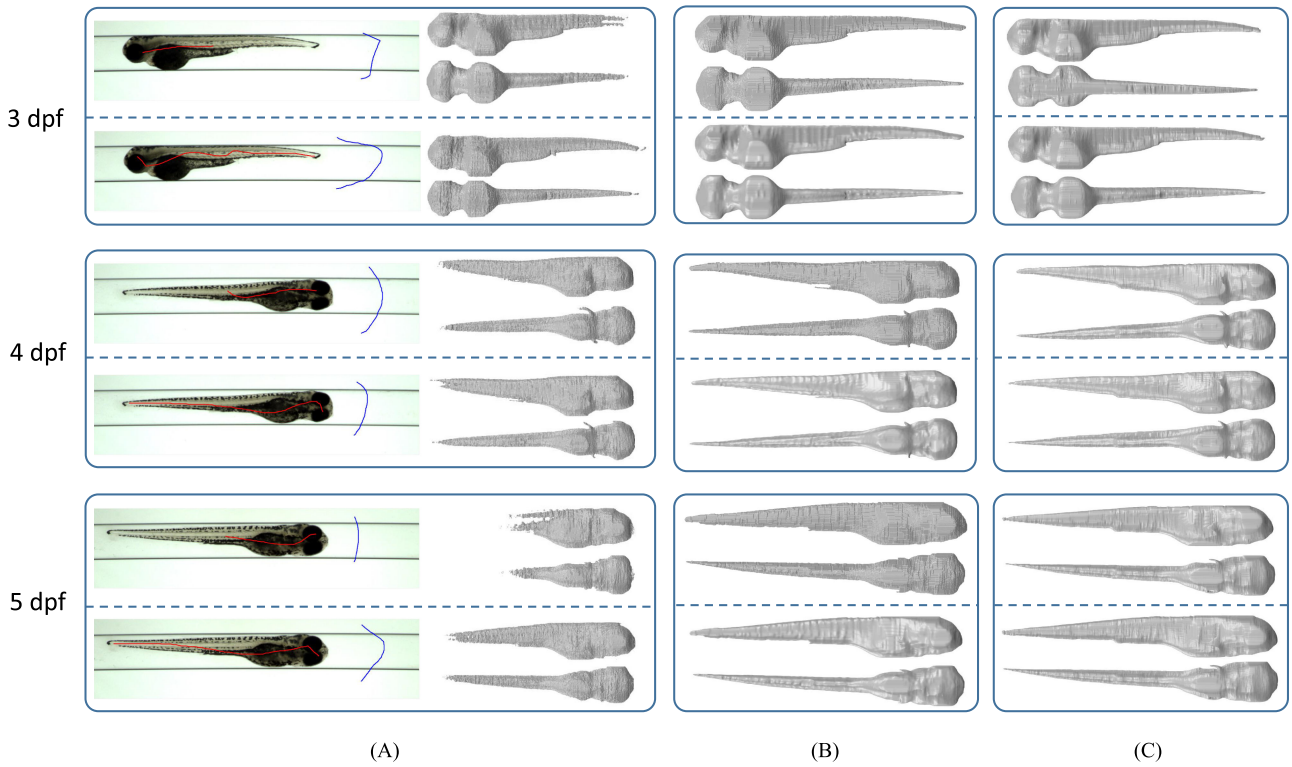


Fig. 3. Comparison of visualised results of different methods. We select one example from each larval stage, i.e., 3, 4 and 5 dpf. For each 3D model, we visualize the same perspectives, i.e., lateral and ventral. (A) The 3D reconstruction of the interactive method. The users randomly draw on the object of interest (red scribble) and background (blue scribble), respectively. Two types of user input are considered, as shown on the left. Corresponding 3D reconstructed models are shown on the right. (B) The 3D reconstruction of the shape-based 3D reconstruction. The raw 3D models are shown in the top two rows. The 3D models with extra surface refinement are shown in below. (C) The 3D reconstructions of the 2-3DLA using only the original-scale textures and multi-scale textures are separately shown on the above and below.

of the specimen become more separable from the background. We threshold the texture-augmented images to obtain the 2D shape approximations.

Camera configuration parameterization: The goal of the camera system calibration is to estimate the intrinsic and extrinsic camera configuration. The intrinsic configuration refers to the innate characteristics of a camera, i.e., focal length of the camera lens f, scaling factor (k_x, k_y) and centroid of the imaging sensor (c_x, c_y). The camera intrinsic configuration is usually provided by the manufacturer and remains stable during imaging. The extrinsic parameters are the transformations from the center of the camera lens to the origin of the 3D world frame. We represent both types of camera configurations as a vector $\psi = (f, k_x, k_y, c_x, c_y, \mathbf{T}, \Theta)^T$, where $\mathbf{T} = (t_x, t_y, t_z)^T$ and $\Theta = (\theta_x, \theta_y, \theta_z)^T$ denote the 3D translation and rotation of the camera center w.r.t. the 3D world frame of reference. In our imaging architecture, the camera position is fixed. The axial-views are generated from the rotation of the specimen along the profile-axis. In order to make the camera configuration parameterizable, we arrange the camera at evenly sampled locations in a (virtual) circle around the specimen as shown in Fig. 1(B). So, we can obtain an extended camera configuration $\psi = (f, k_x, k_y, c_x, c_y, \mathbf{T}_{1:N}, \Theta_{1:N})^T$, where N represents the number of axial-views.

Pinhole camera projection model: For a voxel indexed as $\mathbf{X}_j \in \mathcal{X}\,(j \in [1, M])$, we can obtain its pixel location in an image plane through the pinhole camera projection model $\mathbf{x}_{j\to i} = \mathbf{P}_i \mathbf{X}_j$, where $\mathbf{P}_i^{3\times 4}\,(i \in [1, N])$ is defined as the camera projection matrix for the i_{th} axial-view. Given a specific camera configuration, we can deduce the corresponding $\mathbf{P}_{1:N}$. We first introduce a decomposition for the projection matrix $\mathbf{P}_i = \mathbf{K}\mathbf{R}_i\,(\mathbf{I}|\mathbf{T}_i)$, where,

$$\mathbf{K} = \begin{bmatrix} f * k_x & 0 & c_x \\ 0 & f * k_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

is the intrinsic camera configuration which is shared by the N axial-views; $\mathbf{R}_i$ is the standard 3D rotation matrix derived from the i_{th} camera extrinsic configuration of Θ_i; $\mathbf{I}$ is a 3×3 identity matrix; and $\mathbf{T}_i$ represents the original 3D translation.

Camera system calibration using VRV: For each voxel in the voxel set $\mathcal{X}$, we can determine if it represents an element consisting our desired object according to its visibility w.r.t all axial-views. This can be done through the projections of a voxel onto each 2D shape approximation $\mathbf{S}_i$. If the pixel location of the voxel is covered by all the 2D shape approximations, we keep this voxel as an element from our object. We first formulate this process to run through all the voxels for an axial-view:

$$\mathcal{X}_i^* = \{\mathbf{X} | \mathbf{S}_i(\mathbf{x}_{j\to i}) \neq 0;\, \mathbf{x}_{j\to i} = \mathbf{P}_i(\psi)\mathbf{X}_j,\, \mathbf{X}_j \in \mathcal{X}\} \tag{2}$$

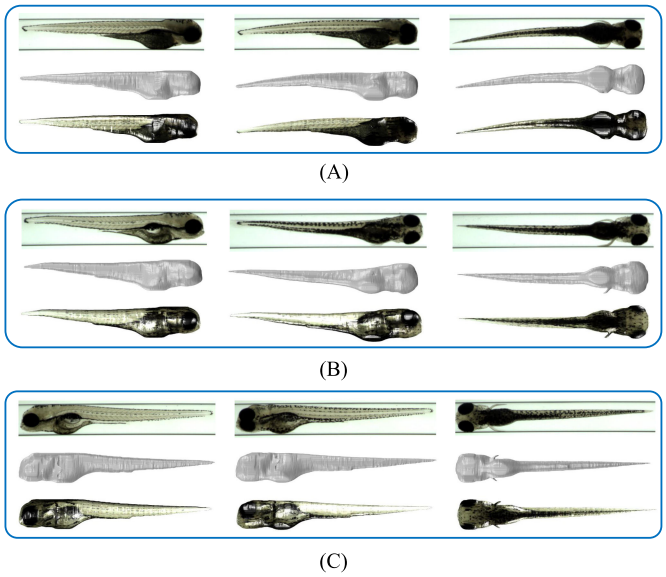


Fig. 4. 3D models visualization of the zebrafish in various larval stages from *Dataset A*. Each bounding box denotes a staged zebrafish. For each specimen, three typical perspectives are shown in different columns. The original axial-view images are shown in the top row; the pure 3D models are shown in the middle row; the texturized models are shown in the bottom row.

Where $\mathcal{X}^*$ represents the subset of the voxels which are visible to the i_{th} axial-view. We re-formalize the projection matrix as $\mathbf{P}_i(\psi)$ to indicate its functional relationship to the camera configuration. When introducing all the N axial-views, we can compute the intersection of the voxel subsets and obtain the elements of our object $\mathcal{X}^* = \bigcap_{i=1:N} \mathcal{X}_i^*$.

We define that the optimal camera configuration will generate an object with the maximal volume of the voxel subset, thus:

$$\psi^* = \arg\max_{\psi \in \Psi} \{|\mathcal{X}^*|\} \tag{3}$$

Where $|\mathcal{X}^*|$ is the defined voxel residual volume which is, in fact, the total voxel number of the voxel subset. For this unconstrained nonlinear optimization problem, we can employ the conventional Nelder-Mead simplex method [46] or an evolution strategy [47]. However, the optimization process may be trapped in many local minima. In practice, we can solve this problem by estimating a good initialization for the camera configuration using the initial setup of our imaging architecture, for example, the 3D rotation angle $\theta_{i,x}$ can be estimated as $\theta_{i,x} = (2\pi) * (i-1)/N$. One should note that the initial camera configuration estimation is far from accurate due to the misalignment of the object center to the camera center and mechanical drift.

B. 3D Volumetric Representation in Light Microscopy Axial-View Imaging

With the established camera configuration, we can estimate a 3D volumetric representation of the object. The quality of the 3D representation for the object can result in various 3D reconstruction effects. The 3D volumetric representation of the specimen either takes a binary value or a probability as the entry for each voxel. The former can offer a compact 3D shape description of the object, but lacks flexibility in describing the surface variation. The latter affords more exploration for the 3D reconstruction. In this work, the binary volumetric representation obtained from the shape-based 3D reconstruction is used as the "groundtruth".

1) 3D Volumetric Representation as Binary Shape: We have obtained the 3D shape of the object as a binary volumetric representation by shape-based 3D reconstruction [10]. This offers an accurate shape reference for the specimen. The method is implemented by the intersection of the cone-shaped camera projections imposed from 2D shapes as obtained from segmentation. The voxels which are included by the intersection of all projections are preserved and used to represent the visible elements from the specimen. Instead of wiping out the invisible voxels to any of the axial-views, a confidence score is assigned to each voxel indicating its visibility to all the views. This is defined as:

$$\mathbf{V}_j = \sum_{i=1:N} \mathbf{1}\left[\mathbf{S}_i(\mathbf{x}_{j\to i}) \neq 0\right] \tag{4}$$

where $\mathbf{1}[\bullet]$ denotes an indicator function which takes the value of one for the voxels that satisfy the condition and zero otherwise. This definition is to an extent less subjected to segmentation errors in 2D shapes. When a threshold $THR \in [1, N]$ is provided, the marching cubes algorithm [48] can evaluate an enclosed surface in this volumetric representation.

2) Improved 3D Volumetric Representation as Confidence Map: In shape-based 3D reconstruction we need accurate segmentations of the original images. These are, however, not always satisfactory. One of the important complications is the translucency and transparency of the specimen in the images. We, therefore, propose the improved 3D volumetric representation in the form of a confidence map which is derived from the probabilistic models as estimated from the object and background presented in the axial-view images.

We can observe that the partial transparency of the specimen is difficult to recover. Severe texture variation within the object challenges the generality of the probabilistic models. We also need to collect prior knowledge on the approximate locations of the objects of interest and background. This can be implemented in an interactive manner. In this case, the results of the 3D reconstruction will rely on the quality of the user input; insufficient user input tends to produce an underestimated model. In Section IV-A, we have proposed to apply the MS algorithm to obtain the approximations of the 2D shapes for the specimen. This also results in an augmentation of the texture representation in the transparent parts of the specimen. The approximations of the 2D shapes can roughly separate the object and background. Although such shape approximations are inaccurate, morphological operations to the envelope of the shape (dilation and erosion) enforce the discrimination of the object and background. In this manner, we have solved the problem of texture augmentation for the partial transparent specimens. In addition, the obtained 2D shape approximations can serve as the shape constraints, such that almost all texture information included by the object can be sampled. This improves the

generalization of the probabilistic models and contributes to the successful automation of the whole system.

We combine the multi-scale textures interpreted as image pyramids to further enhance the probabilistic models. Now, let $\mathcal{I}_{sc} = \{\mathbf{I}_{(1,sc)}, \mathbf{I}_{(2,sc)}, ..., \mathbf{I}_{(N,sc)}\}$ and $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_N\}$ denote the images (after mean shift filtering) in scale of sc and the approximations of 2D shapes in original scale, where N represents the number of views. In any of the views, the foreground and background can be approximately discriminated and indexed by $\mathbf{x}_f$ and $\mathbf{x}_b$, where $\mathbf{S}(\Omega) \approx \{\mathbf{x}_f, \mathbf{x}_b\}$ and Ω represents the whole image domain. Then, the multi-scale textures are represented as $\mathbf{I}_{sc}(\mathbf{x}_f)$ and $\mathbf{I}_{sc}(\mathbf{x}_b)$, where $\mathbf{I}_{sc}(\mathbf{x}) = (r,\ g,\ b)^T$ denote the RGB values in color space for any pixel in scale sc. Next, a multivariate Gaussian is applied to estimate the probabilistic models of the range images for each specific scale. Thus:

$$p_{(f,sc)}(\mathbf{x}_f) = \frac{1}{\sqrt{(2\pi)^3|\mathbf{\Sigma}_f|}} e^{-\frac{1}{2}(\mathbf{I}_{sc}(\mathbf{x}_f)-\mu_f)^T \mathbf{\Sigma}_\mathbf{f}^{-1}(\mathbf{I}_{sc}(\mathbf{x}_f)-\mu_f)}$$
$$p_{(b,sc)}(\mathbf{x}_b) = \frac{1}{\sqrt{(2\pi)^3|\mathbf{\Sigma}_b|}} e^{-\frac{1}{2}(\mathbf{I}_{sc}(\mathbf{x}_b)-\mu_b)^T \mathbf{\Sigma}_\mathbf{b}^{-1}(\mathbf{I}_{sc}(\mathbf{x}_b)-\mu_b)} \tag{5}$$

where $p_{(\bullet,sc)}$ represents the probabilistic model estimated for the textures in scale sc; and $(\mu_f,\ \mathbf{\Sigma}_f)$ and $(\mu_b,\ \mathbf{\Sigma}_b)$ denote the mean vectors and the covariance matrices estimated for the foreground and background, respectively. To avoid overestimation of the probabilistic models, we randomly select a subset of the axial-views for computation.

We have already learned (cf. Section IV-A) that a voxel $\mathbf{X} \in \mathcal{X}$ in 3D space corresponds to a pixel location in $\mathbf{I}_{(i,sc)}$ (the i^{th} image in scale sc) through the pinhole camera model $\mathbf{x} = \mathbf{P}_i\mathbf{X}$, where $\mathbf{P}_i \in \mathcal{P}$ is the camera projection matrix estimated for each view. We can obtain the probabilities of this voxel being the foreground and background by $p_f(\mathbf{x})$ and $p_b(\mathbf{x})$. Suppose that all the N axial-view images $\mathcal{I}_{sc}$ in scale sc are independent, the joint probabilities of a voxel $\mathbf{X}$ indicating its visibility to the foreground and background can be modelled as follows:

$$P_{(f,sc)}(\mathbf{X}) = \left(\prod_{i=1:N} p_{(f,sc)}(\mathbf{x})\right)^{\frac{1}{N}}$$
$$P_{(b,sc)}(\mathbf{X}) = 1 - \left(\prod_{i=1:N} (1 - p_{(b,sc)}(\mathbf{x})\right)^{\frac{1}{N}} \tag{6}$$

We can interpret (6) as the intersection of the N camera projections for each voxel. We use the logarithm to re-scale the two probabilities. The multi-scale probabilistic models are then fused to obtain the confidence map defined as:

$$\mathbf{I}^* = \sum_{sc=1}^{M} \beta_{sc} \left[\log(P_{(f,sc)}) - log(P_{(b,sc)})\right] \tag{7}$$

where $\beta = \{\beta_{sc}\}_{sc=1:M}$ specifies the weight for each scale and $\mathbf{1}^T \cdot \beta = 1$. In our specific implementation, a uniform distribution of β is used, which simply averages the multi-scale probabilities. We use three-scale textures ($M = 3$) in a hierarchical fashion, such that the higher scales can be generated by halving the images in respective lower scales. The confidence map takes a confidence score as the entry for each voxel to indicate its likelihood to be the object (or background). The larger the confidence score of a voxel is, the more possible that voxel is classified as the object.

C. 3D Reconstruction as Objective Function Optimization

With the confidence map, we can accomplish the 3D reconstruction by searching for an enclosed surface which is able to correctly classify all the voxels into the object and background. The marching cubes algorithm [48] straightforwardly estimates a specific surface. However, it is difficult to determine which is the optimal one; and serious fractal behaviour of the estimated surface will occur. This will result in inaccurate 3D measurements of volume and surface area. An empirical solution can be applied to cascade a surface refinement module which needs to be carefully validated. In this work, we apply the region based level set method (CV model) [49] to accomplish this task. The employment of the CV model transfers the 3D reconstruction problem into a 3D segmentation. The optimal enclosed surface retaining a smooth appearance is achieved.

According to the CV model, an enclosed 3D surface $\mathcal{C}$ is embedded in a distance regularized level set function (LSF) $\mathcal{C} := \{\mathbf{X} \in \Omega; \phi(\mathbf{X}) = c\}$, where Ω and $\mathbf{X}$ are consistently defined as the image domain (3D in this case) and the spatial locations of the voxels in world frame. To separate the object and background according to the similarity in the confidence map we define the objective function (8). This includes the external force as the confidence map and the internal force as the smooth appearance of the surface. Thus:

$$E(\phi) = \underbrace{\mu \int_\Omega |\nabla H(\phi)| d\mathbf{X}}_{\text{internal force}} + \underbrace{\upsilon \int_\Omega \left[|\mathbf{I}^* - u_{in}|^2 H(\phi) + |\mathbf{I}^* - u_{out}|^2 (1 - H(\phi))\right] d\mathbf{X}}_{\text{external forece}} \tag{8}$$

where, u_{in} and u_{out} represent the mean intensity of the confidence map in and outside of the surface; μ and υ are constants used to balance different terms. H is the Heaviside function defined as follows:

$$H_\epsilon(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \tag{9}$$

Using the Euler-Lagrange equation, as:

$$\frac{\partial \phi}{\partial t} = -\frac{\partial E}{\partial \phi}, \tag{10}$$

one can implement the surface evolution by computing the gradient flow for the objective function defined as:

$$\frac{\partial \phi}{\partial t} = \delta(\phi)\left\{\mu\, div\left(\frac{\Delta\phi}{|\Delta\phi|}\right) + \upsilon\left[\left(|\mathbf{I}^* - u_{out}|^2 - |\mathbf{I}^* - u_{in}|^2\right)\right]\right\}, \tag{11}$$

where $\delta(\phi)$ is the derivative of $H(\phi)$. By setting a proper step size Δt, we can complete the 3D reconstruction via gradient descent, defined as:.

$$\phi_{t+1} = \phi_t + \Delta t \frac{\partial \phi}{\partial t} \tag{12}$$

3D measurements inference: Given the 3D reconstruction represented as the optimal surface embedded in ϕ, we can derive the 3D measurements. The volume is obtained by the integration over all the voxels which are included in the object ϕ: $V = \sum_{\phi} \mathbf{1}[\phi > c^*] \times v$, where $\mathbf{1}[\bullet]$ keeps its representation as an indicator function; c^* is the optimal c-level set which we will investigate in following sections; v is the unit volume for the voxels which is pre-defined. A set of surface points are generated from the the c^*-level set of ϕ, which will be used to produce a triangulated mesh. The surface area is obtained by the integration over all the facets in the triangulated meshes using Heron's formula $A = (s(s-a)(s-b)(s-c))^{1/2}$, where $s = (a+b+c)/2$ and a, b, c represent the edges of a triangle.

V. Experiments

In this section, we describe our experiments to evaluate the performance of the proposed approach for our datasets. We also present and carefully analyse the experimental results.

A. Visual Inspection of the Results on Dataset A

We first apply different methods on the *Dataset A* to compare the obtained 3D models through visualization. We next apply our 2-3DLA approach on the whole *Dataset A* to demonstrate its performance on the zebrafish in various developmental stages.

1) Comparison With Various Methods Using Visualization: We implemented an interactive method [35] which requires user input to indicate the approximate locations of the object and background. So, the quality of the 3D models will rely on the user input. The method presented in Section IV-B1 is also used in this comparison.

In Fig. 3(A), the results obtained by the interactive method are shown. Two types of user input are employed to estimate the probabilistic models for each specimen. One input collects less textures by drawing shorter scribbles on the object. The other input covers a larger region of the object, thus collecting more texture information. In Fig 3(A), the corresponding 3D reconstructions are shown next to the input drawings. We are confirmed in our assumption that variations in user input impact the quality of the 3D reconstructions. Sufficient texture sampling usually results in better 3D reconstructions preserving most of the shape of the specimen. However, even with "good" user input, parts of the zebrafish, mostly at the tail area, still fail to be faithfully reconstructed. The main reason is the considerable texture inhomogeneity of the specimen which hampers the generality of the probabilistic models and thereby results in a misclassification of the translucent part of the specimen as background. From the results of the interactive method, we can also observe the fractal behaviour of the surface. It is apparent that inaccurate 3D reconstruction will result in inaccurate 3D measurements.

In Fig. 3(B) we show results of the shape-based method with a set of accurate 2D shapes, which generates very accurate 3D shape representation of the specimen. Each shape shows a natural and detailed surface, but fractal behaviour also occurs due to the carving effects of the method. This is depicted in the top two rows of Fig. 3(B). After extra surface refinement [50], accurate 3D measurements can be available. The refined 3D models are shown in the bottom two rows Fig. 3(B). We need to accept that the shape-based method requires accurate 2D segmentations from the original images which sometimes are difficult to obtain in light microscopy imaging.

In Fig. 3(C) we show the 3D models obtained by the 2-3DLA system. In the top two rows, the 3D reconstructed models are obtained by the confidence map derived with one-scale texture. In the bottom two rows, the results obtained with three-scales texture are shown. One can observe a better 3D visualization for both of the results. In the 2-3DLA, the employment of the MS algorithm ensures improved discrimination between the object and background. Furthermore, the shape constrained texture collection augments the generality of the confidence map. These improvements cooperate in obtaining a better volumetric representation. The CV model aims to search for the optimal surface which can largely separate the object and background in the confidence map and preserves a smooth appearance of the specimen, which subsequently results in accurate 3D measurements. However, a little fractal behaviour is still showing in the results with one-scale texture, resulting in a slightly deficient tail region of the zebrafish.

2) 3D Visualization for Dataset A Using 2-3DLA: Here we apply the 2-3DLA on the whole *Dataset A*. In Fig. 4 some results are depicted. For each developmental stage of the zebrafish, we randomly select one specimen in the dataset. To create additional value for 3D models, we map the textures from the original images onto the surface of the 3D models to produce texturized 3D shapes. For practical reasons, the texturing of the partial transparent regions is artificial. We separately show the binary and the texturized 3D models in the middle and bottom rows for each specimen. We select three typical axial-views, i.e., lateral, tilted and ventral.

From the visualization, we can assess that the 2-3DLA obtains very natural and vivid 3D shapes for the specimens. The problem arising from the translucency and transparency seems to be solved. From the 3D shapes it is obvious that older zebrafish are larger compared to the younger ones. The size of yolk in the various stages is different. During the development, the yolk of the zebrafish provides nutrient for its growth. Consequently, older zebrafish have smaller yolk, simply as a result of their growth. Furthermore, the 3D shape obtained by the our approach offers a good framework for the visualization of the biological system

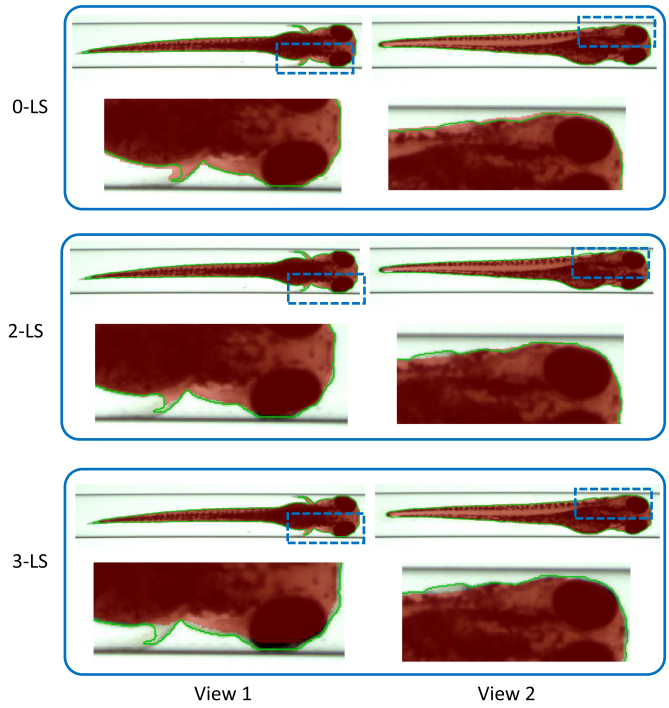


Fig. 5. 2D shape coherence (area in red) between the projections from 3D models to axial-view images and the groundtruth 2D shapes (green contours) with various c-level sets. A subsection of the original 2D shape coherence (dashed bounding box) is zoomed in and shown at the bottom of each row.

by introducing organ-level 3D systems, e.g., the circulatory system, the liver, the brain etc. Using specific staining these can be incorporated and visualized. The whole process is reviewed in a 2-3DLA spotlight including animated visualizations of the 3D models at http://bio-imaging.liacs.nl/galleries/VAST-2-3DLA/.

B. 3D Measurements for Dataset A

We investigate how to obtain accurate 3D measurements from the 3D models obtained by the 2-3DLA. To this end, the hyper-parameter of c-level set needs to be tuned, because an unvalidated c-level set may lead to inaccurate 3D measurements.

1) 2D Shape Coherence: A good 3D reconstruction must approximate the real shape of the object. We evaluate and interpret this as shape coherence. We define the 2D shape coherence as a disparity map from the projected shape of a 3D model to a "groundtruth 2D shape" in each original axial-view. We have developed a robust method for the segmentation of zebrafish larvae. Using manual segmentations, the method is validated and turned out to be accurate. We thereby use the segmentation results as the approximation of the groundtruth 2D shape for the *Dataset A*. Accordingly, we interpret the 2D shape coherence as the overlap from the projected shape to the groundtruth 2D shape. We chose three different c-level sets, i.e., 0, 2 and 3, to extract the corresponding 3D models. The 2D shape coherence that are obtained in this manner are shown in Fig. 5. We depict three typical axial-views of the object, i.e., ventral and lateral. We represent the groundtruth 2D shapes as green contours. In order to appreciate the results, a subsection of the 2D shape coherence is zoomed in and its visualization is shown in the bottom row for each view.

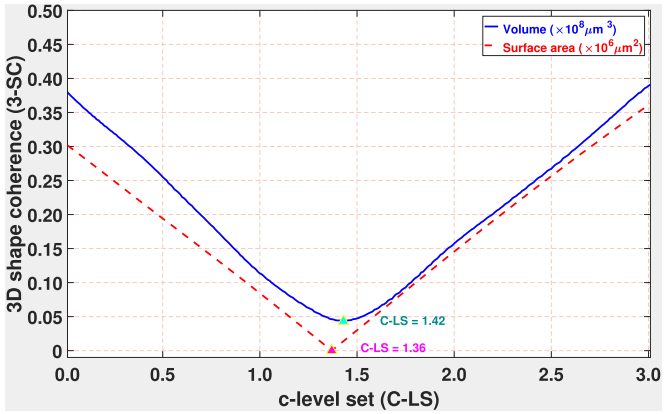


Fig. 6. 3D shape coherence with various c-level sets. The triangles indicate the minimum of the measurements, which are found around 1.4-level set. To ensure the correct visualization, we normalized both measurements.

One can observe that a small c-level set, i.e., 0-level set, produces an overestimation of the 3D model, which presents a large 2D shape coherence. This is illustrated as the red area exceeds the green contours in the top row of Fig. 5. A large c-level set, i.e., 3-level set, results in a more compact 3D shape of the object, which obtains a smaller 2D shape coherence. This is illustrated as the red area separating from the green contours in the bottom row of Fig. 5. In contrast, a reasonable c-level set, i.e., 2-level set, obtains highly correlated 2D shape coherence producing a better 3D reconstruction.

2) 3D Shape Coherence: The shape-based 3D reconstruction method results in a binary volumetric representation. If accurate 2D shapes are used in this method, the obtained 3D model can serve as the "groundtruth". The 3D shape coherence is then defined as the disparity from the 3D models obtained by the 2-3DLA to the "groundtruth 3D model". We propose two measurements to evaluate the 3D shape coherence: volume disparity ($\overline{V}$) and surface area disparity ($\overline{A}$). They are separately defined as $\overline{V} = [(1/M)\sum_i (V_i^* - V_i^m)^2]^{(1/2)}$ and $\overline{A} = [(1/M)\sum_i (A_i^* - A_i^m)^2]^{(1/2)}$, where V^* and A^* are the volume and surface area measured from the 3D model groundtruth; V^m and A^m are the volume and surface measured from the 3D models obtained using a specific c-level set; and M is the number of specimens. We use half of the *Dataset A* for the validation of this experiment. In Fig. 6, we show a graph of the 3D shape coherence for various c-level sets. We illustrate that the choice of a small c-level set, e.g., 0-level set, tends to produce an overestimation of the measurements. The choice of a large c-level set, e.g., 2-level set, will lead to an underestimation. This corresponds nicely with the findings in the 2D shape coherence as shown in Fig. 5. Finally we find that the 1.4-level set is a good compromise and gives the best estimation of the volume and surface area. For the next section, we use the 1.4-level set to obtain the deterministic 3D models from which we can derive the 3D measurements.

3) 3D Measurements for Dataset A: We use the remaining half of the *Dataset A* to compute the 3D measurements of the

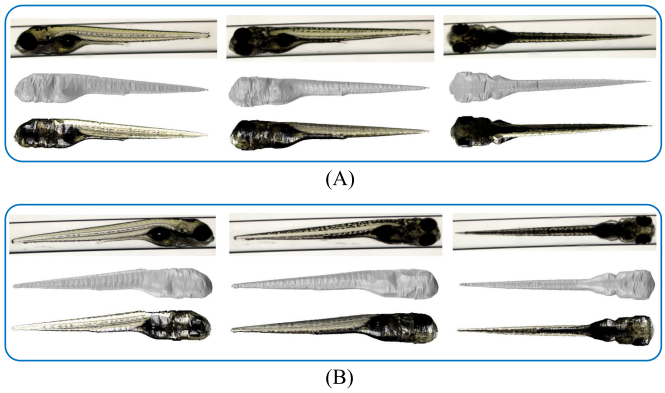


Fig. 7. Visualization of 3D models of the zebrafish in bright-field microscopy from *Dataset B*. For each specimen, three typical axial-view images are shown on the top; the corresponding axial-views of the pure and texturized 3D models are shown in the middle and bottom, respectively.

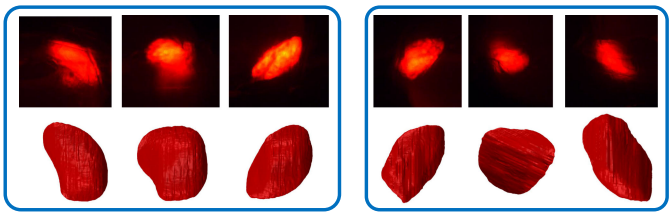


Fig. 8. Visualization of 3D models of the zebrafish liver in fluorescent microscopy from *Dataset C*. We show three axial-view images and the corresponding 3D perspectives in the top and bottom row for each specimen.

TABLE I
VOLUME STATISTICS ($\times 10^8\,\mu$M^3) OF THE 3D MODELS FOR THE ZEBRAFISH IN VARIOUS DEVELOPMENTAL STAGES

	Shape	2-3DLA
3 dpf	2.53 ± 0.11	2.56 ± 0.12
4 dpf	2.63 ± 0.19	2.66 ± 0.19
5 dpf	3.00 ± 0.18	2.96 ± 0.18

• Shape = Shape-based 3D reconstruction with surface refinement • 2-3DLA = The proposed approach

TABLE II
SURFACE AREA STATISTICS ($\times 10^6\,\mu$M^2) OF THE 3D MODELS FOR THE ZEBRAFISH IN VARIOUS DEVELOPMENTAL STAGES

	Shape	2-3DLA
3 dpf	3.20 ± 0.15	3.21 ± 0.14
4 dpf	3.34 ± 0.17	3.36 ± 0.17
5 dpf	3.63 ± 0.14	3.61 ± 0.14

zebrafish in various developmental stages, i.e., 3, 4 and 5 dpf. In this experiment, we compute the statistics for the volume and surface area of the zebrafish and show the results in Tables I and II. The 3D measurements obtained from the shape-based 3D reconstruction method are used to assess the performance of the 2-3DLA.

TABLE III
VOLUME ($\times 10^8\,\mu$M^3) AND SURFACE AREA ($\times 10^6\,\mu$M^2) OF THE 3D MODELS FOR THE ZEBRAFISH IN *Dataset B*

	Volume		Surface area	
	Shape	2-3DLA	Shape	2-3DLA
#1	3.00	3.05	3.61	3.65
#2	2.95	3.04	3.52	3.60
#3	3.06	3.13	3.69	3.74

TABLE IV
VOLUME ($\times 10^5\,\mu$M^3) AND SURFACE AREA ($\times 10^4\,\mu$M^2) OF THE 3D MODELS FOR ZEBRAFISH LIVER IN *Dataset C*

	Volume		Surface area	
	Shape	2-3DLA	Shape	2-3DLA
#1	7.70	7.56	4.67	4.57
#2	5.38	4.15	3.61	3.15
#3	8.10	8.47	4.91	5.04
#4	9.06	9.28	5.20	5.18
#5	11.6	9.77	6.07	5.49
#6	15.1	16.6	6.89	7.69
#7	6.55	6.79	4.28	4.31

From the tables one can conclude that the 2-3DLA can obtain accurate 3D measurements of volume and surface area comparable to the groundtruth method. This holds for the zebrafish in the three developmental stages. The results of the 2-3DLA are slightly larger, but the difference is small and acceptable. One should note that the 2-3DLA will be advantageous for the cases in which accurate segmentations are not available. Furthermore, the increasing trends of the volume and surface area for the zebrafish staged from 3 dpf to 5 dpf is consistent with the expected physical growth of the zebrafish. Importantly, Tables I and II offer baseline metrics for volume and surface area of the zebrafish in various stages; this facilitates quantitative and qualitative analysis using the zebrafish as model system.

C. 3D Visualization and Measurements for Dataset B & C

We have shown that the 2-3DLA can obtain an accurate 3D shape description of the whole specimen offering the baseline reference for the 3D measurements. This is, however, not sufficient to evaluate a specimen on the level of organs. From the optics used, the resolving power of *Dataset B* & *C* (cf. Section III) is much higher. Moreover, different microscope modalities are used for these datasets, i.e., bright-field and fluorescence. As a case study, in addition to the whole specimen, we use the zebrafish liver as a model organ system. The liver size is important to pharmacokinetics, as it is an crucial organ for drug metabolism. For *Dataset B* & *C*, 3D measurements of volume and surface area are separately reported in Tables III and IV. As a comparison, we still use the results obtained from the shape-based 3D reconstruction approach.

TABLE V
PERFORMANCE EVALUATION ON RUNTIME (S=SECOND,MIN=MINUTE)

	Step A	Step B	Step C	Step D	Total
Dataset A	36 ± 1.4(s)	12.6 ± 0.8(s)	35.9 ± 0.9(s)	2.2 ± 0.5(min)	~ 3.6(min)
Dataset B	4.3 ± 0.2(min)	56.4 ± 0.9(s)	37.2 ± 0.4(min)	7.9 ± 0.5(min)	~ 50.3(min)
Dataset C	5.5 ± 0.6(min)	39.0 ± 7.4(s)	8.8 ± 1.2(min)	12.1 ± 1.1(min)	~ 27.1(min)

• Step A: Image capture • Step B: 2D shape approximation • Step C: Camera system calibration • Step D: 3D reconstruction

In Table III we observe only a small difference of the volume and surface area produced by the 2-3DLA compared to baseline method. This difference is acceptable. It is apparent that both of the measurements are within the distributions given in Tables I and II. This suggests that the 2-3DLA system shows stable performance for bright-field microscopy axial-view imaging. Subsequently, it can obtain accurate 3D measurements for the specimen as a whole. In Fig. 7, we show the 3D models for two specimens. These results are presented in the similar fashion as used in Fig. 4. An interesting phenomenon shown in the Example #1 is that the tail of the 3D model is thinner than the original specimen. This is because, in this experiment, the specimen were alive and anesthetized. During the imaging, the motion of the positioning capillary introduces slight shape deformations of the specimen. This impedes accurate 3D reconstruction which requires static objects.

In Table IV we can find that most of the zebrafish livers are reconstructed well and comparable 3D measurements are obtained. Significant difference for the 3D measurements are found in example #2. Both the volume and surface are obtained by the 2-3DLA are smaller than that obtained by the baseline method. This is caused by the image quality for the example #2. We visualize the 3D model for this example in Fig. 8(B). One can observe that the original axial-view images do not clearly represent the specimen. We used a relatively small external force in (10) for all the specimens, which produces compact 3D models. In fact, using a different configuration for the parameter in the example #2 will result in accurate 3D measurements. However, in order to ensure a justified evaluation of this experiment, we reported the results obtained with the same parameters. In future work, we can validate the parameter specifications in relation to image quality.

D. Evaluation on Efficiency

For the evaluation of the computational efficiency of our approach, we first specify the setup. We represent the process of image capture, 2D shape approximation, camera system calibration and the 3D reconstruction as Step A, B, C and D, respectively. In Step A, we use the VAST-BioImager to automatically obtain *Dataset A*. A well-trained biologist used the microscope setup (Leica DMRB) to acquire *Dataset B* & *C*. In this data acquisition, the object is manipulated by the VAST-BioImager to present in the view of the microscope. The remaining steps are implemented with the Matlab platform (and partial C/C++ implementation) on a desktop equipped with a CPU i7 and 16G RAM. In Step B, we heuristically configure the filter width in spatial, color feature space as $(h_s, h_r) = (10, 20)$ for *Dataset A*; and $(h_s, h_r) = (20, 30)$ for *Dataset B* & *C*. In Step C, we find that the Nelder-Mead simplex method works much faster for *Dataset A* & *C*; The evolution strategy achieves robust performance for *Dataset B*, though it is not efficient. We take a compromise on the optimizers and we apply the Nelder-Mead simplex method for *Dataset A* & *C* and the evolution strategy for *Dataset B*. In phase II of our 2-3DLA, we set the parameters $\mu = 10$, $\upsilon = 1$ and $\Delta t = 0.5$ for all cases.

In Table V, for each step from A to D, we report the performance evaluation in terms of runtime in minutes or seconds for the three datasets. We compute the average and standard deviation of the runtime for each specimen in the datasets. First, from the results, we may conclude that the performance of our approach on *Dataset A* is most efficient. In contrast, the larger size of the images in the *Dataset B* and *C* results in an expensive computation for our approach. Second, Step B performs very efficiently in the three datasets due to our fast implementation of the MS algorithm. Third, Step C in *Dataset B* and *C* requires a large amount of computations. The reason is that during the camera system optimisation, massive evaluations for the cost function are executed. However, a better optimiser and a high-performance scheme such as parallelisation or GPU programming can be employed to improve the efficiency of this step. Fourth, Step D in the three datasets seems be computationally expensive. One reason is the iterative 3D surface evolution. We set up a restricted termination criterion that requires the change of the cost function to be close to zero. So, this results in more iterations. In addition, we set a small grid size for the voxel space. This leads to massive projections and demands a large amount of memory to fulfil a precision guarantee. Again, this can, to a certain extent, be solved by the introduction of high-performance computing strategies.

VI. CONCLUSION AND FUTURE WORK

Imaging of partially transparent objects impedes the visualisation of a continuous surface which renders it difficult to estimate a depth map of the object. To solve this problem, we have developed the 2-3DLA approach for 3D reconstruction and measurement in light microscopy axial-view imaging. In phase I, we estimate an improved 3D volumetric representation as a confidence map from a range of texture-augmented images by applying the MS algorithm on the original images. In phase II, we adapt the region based level set method to estimate the optimal enclosed surface for the object which balances the 3D shape integrity and the smoothness of appearance for the 3D

model. Compared with the shape-based method, the proposed 2-3DLA approach obtains comparable 3D reconstructions and measurements, but it does not require accurate 2D segmentations of the original images. As segmentation of partial transparent objects can be difficult, this system provides a good solution to this shortcoming. Unlike the shape-based 3D reconstruction method, the 2-3DLA does not need extra refinement to obtain a natural and smooth 3D surface. We have successfully applied our approach for zebrafish analysis in multi-modal light microscopy axial-view imaging including bright-field and fluorescence. The obtained results can be directly used for the model system evaluation in the fields of toxicology, infectious diseases and oncology.

The efficiency of the our approach still needs improvement with respect to larger images. A possible solution can be a dynamic programming scheme for high-performance parallelization. In the field of life-science, the zebrafish model is intensively used. Therefore, in our 2-3DLA we can learn a good probabilistic model for the zebrafish via either supervised or semi-supervised manner [51]. Our approach still needs the support from 2D shape approximation in the process of camera system calibration. In future work, we aim for a method which is independent on image segmentation. In addition, other multi-view imaging modes need to be investigated with our system. The axial-view imaging is, however, the most commonly used, so the 2-3DLA can be considered generic for other specimen observations.

Appendix A

The zebrafish used in *Dataset C* are from a transgenic line Tg(lfabp:dsRed; elaA:EGFP) with 2 fluorescent colors for liver and pancreas (2CLIP). The eggs were kept in 60 μg/mL Instant Ocean Sea Salts (Sera Marin, Heinsberg Germany) in demineralized water and treated with 0.003% 1-phenyl-2-thiourea (PTU, Sigma-Aldrich, Zwijndrecht, The Netherlands) to prevent pigmentation.

Acknowledgment

The authors would like to thank Dr. B. Koch from the Department of Animal Sciences and Health, Institute of Biology, Leiden University and R. van Wijk from Leiden Academic Centre for Drug Research, Leiden University for their contribution of experimental datasets.

References

[1] A. Nezhinsky and F. Verbeek, "Pattern recognition for high throughput zebrafish imaging using genetic algorithm optimization," in *Proc. IEEE Conf. Pattern Recognit. BioInformat.*, 2010, pp. 301–312.

[2] V. Kantae *et al.*, "Pharmacokinetic modeling of paracetamol uptake and clearance in zebrafish larvae: Expanding the allometric scale in vertebrates with five orders of magnitude," *Zebrafish*, vol. 13, no. 6, pp. 504–510, 2016.

[3] Z. Xiong and F. Verbeek, "Segmentation of zebrafish larva inhomogeneous 3d images using the level-set method," *Electron. Imag.*, vol. 2016, no. 1, pp. 1–9, 2016.

[4] S. Arridge, "Optical tomography in medical imaging," *Inverse Problems*, vol. 15, no. 2, pp. R41–R93, 1999.

[5] M. Levoy, R. Ng, A. Adams, M. Footer, and M. Horowitz, "Light field microscopy," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 924–934, 2006.

[6] M. Broxton *et al.*, "Wave optics theory and 3-d deconvolution for the light field microscope," *Opt. Express*, vol. 21, no. 21, pp. 25 418–25 439, 2013.

[7] C. Pardo-Martin, T. Chang, B. Koo, C. Gilleland, S. Wasserman, and M. Yanik, "High-throughput in vivo vertebrate screening," *Nature Methods*, vol. 7, no. 8, pp. 634–636, 2010.

[8] A. Fitzgibbon, G. Cross, and A. Zisserman, "Automatic 3d model construction for turn-table sequences," in *Proc. Eur. Workshop 3D Struct. Multiple Images Large-Scale Environ.*, 1998, pp. 155–170.

[9] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[10] Y. Guo, W. Veneman, H. Spaink, and F. Verbeek, "3d zebrafish larvae reconstruction and measurements with high-throughput in vivo axial-view imaging," *Biomed. Opt. Express*, vol. 8, no. 5, pp. 2611–2634, 2017.

[11] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[12] A. Sacco *et al.*, "Cancer cell dissemination and homing to the bone marrow in a zebrafish model," *Cancer Res.*, vol. 76, no. 2, pp. 463–471, 2016.

[13] Z. Zhang, R. Deriche, O. Faugeras, and Q. Luong, "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry," *Artif. Intell.*, vol. 78, no. 1, pp. 87–119, 1995.

[14] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 1, pp. 519–528.

[15] J. Wei, B. Resch, and H. Lensch, "Multi-view depth map estimation with cross-view consistency," in *Proc. Brit. Mach. Vis. Conf.*, 2014. [Online]. Available: https://scholar.google.com/scholar?q=Multi-view+depth+map+estimation+with+cross-view+consistency&btnG=&hl=nl&as_sdt=0%2C5

[16] K. Yoon and I. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 650–656, Apr. 2006.

[17] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3017–3024.

[18] Y. Liu, X. Cao, Q. Dai, and W. Xu, "Continuous depth estimation for multi-view stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2121–2128.

[19] M. Bleyer, C. Rhemann, and C. Rother, "Patchmatch stereo-stereo matching with slanted support windows," in *Proc. Brit. Mach. Vis. Conf.*, 2011, vol. 11, pp. 1–11.

[20] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proc. IEEE Conf. Comput. Vis.*, 2015, pp. 873–881.

[21] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2010.

[22] C. Huang, C. Cagniart, E. Boyer, and S. Ilic, "A Bayesian approach to multi-view 4d modeling," *Int. J. Comput. Vis.*, vol. 116, no. 2, pp. 115–135, 2016.

[23] G. Xu and Z. Zhang, *Epipolar Geometry in Stereo, Motion and Object Recognition: A Unified Approach*, vol. 6. New York, NY, USA: Springer, 2013.

[24] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[25] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to sift or surf," in *Proc. IEEE Conf. Comput. Vis.*, 2011, pp. 2564–2571.

[26] M. Brown and D. Lowe, "Automatic panoramic image stitching using invariant features," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 59–73, 2007.

[27] C. Wu, S. Agarwal, B. Curless, and S. Seitz, "Multicore bundle adjustment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3057–3064.

[28] J. Franco and E. Boyer, "Exact polyhedral visual hulls," in *Proc. Brit. Mach. Vis. Conf.*, 2003, vol. 1, pp. 329–338.

[29] S. Lazebnik, Y. Furukawa, and J. Ponce, "Projective visual hulls," *Int. J. Comput. Vis.*, vol. 74, no. 2, pp. 137–165, 2007.

[30] A. Djelouah, J. Franco, E. Boyer, F. Le Clerc, and P. Pérez, "N-tuple color segmentation for multi-view silhouette extraction," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 818–831.

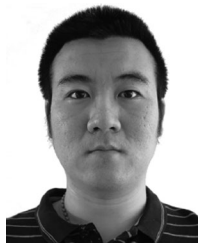
[31] K. Kutulakos and S. Seitz, "A theory of shape by space carving," *Int. J. Comput. Vis.*, vol. 38, no. 3, pp. 199–218, 2000.

[32] Y. Furukawa and J. Ponce, "Carved visual hulls for image-based modeling," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 53–67, 2009.

[33] K. Kolev, M. Klodt, T. Brox, and D. Cremers, "Continuous global optimization in multiview 3d reconstruction," *Int. J. Comput. Vis.*, vol. 84, no. 1, pp. 80–96, 2009.
[34] D. Cremers and K. Kolev, "Multiview stereo and silhouette consistency via convex functionals over convex domains," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1161–1174, Jun. 2011.
[35] K. Kolev, T. Brox, and D. Cremers, "Fast joint estimation of silhouettes and dense 3D geometry from multiple images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 493–505, Mar. 2012.
[36] C. Esteban and F. Schmitt, "Silhouette and stereo fusion for 3D object modeling," *Comput. Vis. Image Understanding*, vol. 96, no. 3, pp. 367–392, 2004.
[37] I. Ihrke, K. Kutulakos, H. Lensch, M. Magnor, and W. Heidrich, "Transparent and specular object reconstruction," *Comput. Graph. Forum*, vol. 29, no. 8, pp. 2400–2426, 2010.
[38] J. Ye, Y. Ji, F. Li, and J. Yu, "Angular domain reconstruction of dynamic 3D fluid surfaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 310–317.
[39] D. Liu, X. Chen, and Y. Yang, "Frequency-based 3D reconstruction of transparent and specular objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 660–667.
[40] Y. Qian, M. Gong, and Y. Yang, "3D reconstruction of transparent objects with position-normal consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4369–4377.
[41] N. Savinov, C. Hane, L. Ladicky, and M. Pollefeys, "Semantic 3D reconstruction with continuous regularization and ray potentials using a visibility consistency constraint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. pp. 5460–5469.
[42] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3279–3286.
[43] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, pp. 1–32, 2016.
[44] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.
[45] C. Hernández, F. Schmitt, and R. Cipolla, "Silhouette coherence for camera calibration under circular motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 343–349, Feb. 2007.
[46] J. Lagarias, J. Reeds, M. Wright, and P. Wright, "Convergence properties of the Nelder–Mead simplex method in low dimensions," *SIAM J. Optim.*, vol. 9, no. 1, pp. 112–147, 1998.
[47] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evol. Comput.*, vol. 9, no. 2, pp. 159–195, 2001.
[48] W. Lorensen and H. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *ACM Siggraph Comput. Graph.*, vol. 21, no. 4, pp. 163–169, 1987.
[49] T. Chan and L. Vese, "Active contours without edges," *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 266–277, Feb. 2001.
[50] M. Desbrun, M. Meyer, P. Schröder, and A. Barr, "Implicit fairing of irregular meshes using diffusion and curvature flow," in *Proc. Conf. Comput. Graph. Interactive Techn.*, 1999, pp. 317–324.
[51] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.



Yuanhao Guo received the M.S. degree in image processing and pattern recognition from Shandong University, Jinan, China, in 2009. He is currently working toward the Ph.D. degree at the Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands. He is working on light microscope imaging and 3-D reconstruction. His interests include applications of machine learning and computer vision.



Yunpeng Zhang received the B.A. degree in microelectronics from Tsinghua University, Beijing, China, in 2012 and the M.S. degree from the Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands, in 2016. His M.Sc. thesis was supervised by Prof. F. J. Verbeek and Dr. Y. Guo.



Fons J. Verbeek received the Ph.D. degree in applied physics (pattern recognition group) from Delft University of Technology, Delft, The Netherlands, in 1995. He is chairing the Imaging and Bioinformatics group, Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands. The research focusses on development of robust methods for large-scale image processing and analysis in the biosciences. He has a particular interest in the zebrafish system for which he is involved in a number of zebrafish projects.

Identifying Surface BRDF From a Single 4-D Light Field Image via Deep Neural Network

Feng Lu, Lei He, Shaodi You, Xiaowu Chen, and Zhixiang Hao

***Abstract*—Bidirectional reflectance distribution function (BRDF) defines how light is reflected at a surface patch to produce the surface appearance, and thus, modeling/recognizing BRDFs is of great importance for various tasks in computer vision and graphics. However, such tasks are usually ill-posed or require heavy labor on image capture from different viewing angles. In this paper, we focus on the problem of remote BRDF type identification, by delivering novel techniques that capture and use a single light field image. The key is that a light field image captures both the spatial and angular information by a single shot, and the angular information enables effective samplings of the four-dimensional (4-D) BRDF. To implement the idea, we propose convolutional neural network based architectures for BRDF identification from a single 4-D light field image. Specifically, a StackNet and an Ang-convNet are introduced. The StackNet stacks the angular information of the light field images in an independent dimension, whereas the Ang-convNet uses angular filters to encode the angular information. In addition, we propose a large light field BRDF dataset containing 47 650 high-quality 4-D light field image patches, with different 3-D shapes, BRDFs, and illuminations. Experimental results show significant accuracy improvement in BRDF identification by using the proposed methods.**

***Index Terms*—BRDF, CNN, light field.**

I. Introduction

LIGHT field imaging captures both the spatial and angular information of a scenario with a single snap shot. By properly balancing the spatial aliasing and angular aliasing, a single light field image can be used to enable various tasks, e.g., depth estimation, refocusing, illumination estimation, material estimation, BRDF (bidirectional reflectance distribution function) estimation, which are not easy for traditional methods with a common image. As a result, researches on light field imaging and analysis are able to benefit various applications of computer vision, computer graphics, virtual reality and augmented reality.

In this paper, we focus on the BRDF identification task. BRDF is the function that describes how a surface reflects light. It determines the surface appearance and thus is of great importance for various tasks in computer vision and graphics. However, modeling/identifying BRDF from a single image is not trivial because a BRDF is a four dimensional function that involves different view angles, light source angles, and surface normal. Therefore, measuring a BRDF typically requires up to hundreds of images [1].

Unlike traditional BRDF estimation methods, which are either ill-posed from a single shot image, or require heavy labor on image capture from different viewing angles [1], we propose a novel method for BRDF identification by using a single light field image. The key is that a light field image captures not only the spatial information, but also the angular information through a single shot. The additional angular information in the 4D light field image provides key information to capture the 4D known BRDF.

In order to fully exploit the information in light field image, this paper proposes convolutional neural network (CNN) based architectures for BRDF identification from a single 4D light field image, together with a new light field BRDF dataset. In particular, we propose two network architectures, namely StackNet and Ang-convNet. The StackNet stacks up the angular information of the light field images in an independent dimension, while the Ang-convNet uses angular filters to directly learn the angular information from each angular block. By doing so, the two methods utilize the angular information in different ways.

As for the dataset, it is produced by using different 3D shapes, BRDFs and illuminations. For the BRDF, we use five representative types: plastic, diffuse, narrow-lobe-metal, wide-lobe-metal, and mid-lobe-metal. Each type contains three sub-types which are similar but not identical. Totally 1200 4D light field images are synthesized, producing 47650 valid light field image patches to form a large light field dataset. To our knowledge, this is the first light field BRDF dataset of this size. The proposed CNNs can be directly trained on this dataset, and accomplish the BRDF identification task. Fig. 1 provides an overview of the proposed architecture.

Manuscript received January 5, 2017; revised April 17, 2017; accepted June 19, 2017. Date of publication July 17, 2017; date of current version October 23, 2017. This work was supported in part by the National Natural Science Foundation of China (No. 61602020), the Joint Funds of NSFC-CARFC (No. U1533129), and the NSFC (No. 61532003, 61325011). The guest editor coordinating the review of this paper and approving it for publication was Dr. Lu Fang. *(Corresponding authors: Feng Lu and Xiaowu Chen.)*

F. Lu is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, and also with the International Research Institute for Multidisciplinary Science, Beihang University, Beijing 100191, China (e-mail: lufeng@buaa.edu.cn).

X. Chen is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: chen@buaa.edu.cn).

L. He and Z. Hao are with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: helei01470@buaa.edu.cn; haozx0@buaa.edu.cn.).

S. You is with the Data61-CSIRO, Canberra, ACT 2601, Australia, and also with Australian National University, Canberra, ACT 0200, Australia (e-mail: shaodi.you@anu.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2728001

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

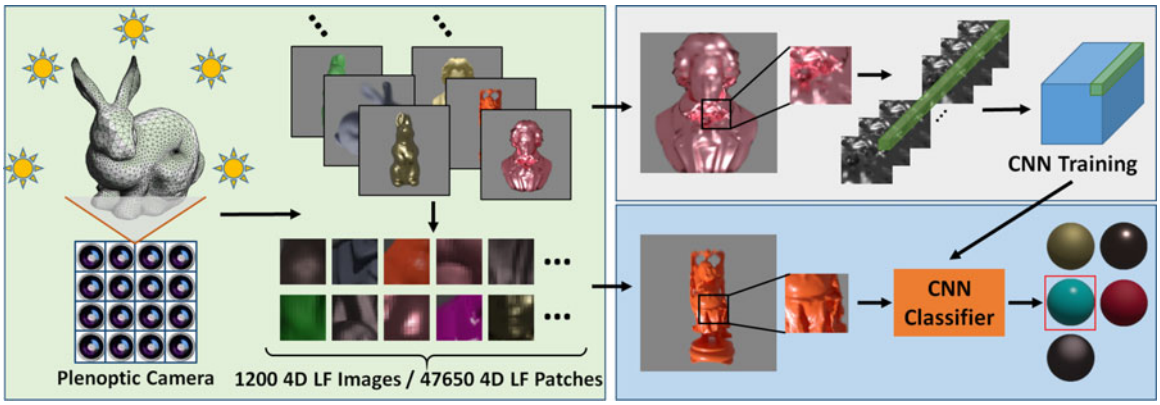


Fig. 1. An overview of the proposed dataset and method for identification of BRDF. Light field images allows simultaneous spatial and angular sampling of an object, which provide rich information for BRDF identification. We collect a dataset containing 47650 4D light field image patches, and train an CNN network with both spatial and angular awareness that automatically identifies the surface BRDF.

The contributions of this paper are summarized as follows. 1) We propose to accomplish the BRDF type identification task by using a single light field image. To the best of our knowledge, this is novel. 2) We propose two CNN based architectures that can effectively utilize light field data for BRDF type identification, which outperform conventional methods by a large margin. 3) We create a large 4D light field dataset that enables efficient training and evaluation.

The rest of the paper is organized as follows: Section II introduces the related works on light field and BRDF. Section III provides the theoretical analysis to show the potential to bridge BRDF identification with light field imaging. Section IV introduces our 4D light field image dataset for BRDF identification followed by the CNN architectures proposed in Section V. Finally, Section VI provides systematical evaluations and Section VII concludes the paper.

II. Related Work

A. Light Field

A light field camera or plenoptic camera allows capturing of spatial and angular information of a scene from a single shot [2]–[7]. This important trait makes light field imaging an interesting topic in computer vision, computer graphics and virtual reality, where the material estimation, illumination estimation, refocusing are essential.

The light field camera was first presented by Lippmann *et al.* [8], and they used lenslets. Ives *et al.*in [9] used pinhole screens. Both of the devices and techniques are based upon the idea that each of the lenslet or pinholes can be considered as a single perspective camera. Because all the lenslets are arranged by varying the spatial positions, they can be viewed as a set of camera array. Stewart *et al.* [10] considered filter to improve sampling aliasing.

A benefit of light field camera is that it avoids angular aliasing as those usually found in camera arrays and multi-view stereo cameras [4]. However, the angular sampling accuracy is achieved through a trade-off with its spatial resolution. Thus, a light field camera usually suffers from spatial aliasing [11], and spatial aliasing is dependent on the camera geometry [3], For a better solution of the trade-off, Levoy and Hanrahan [5] considered the use of large apertures. Bishop *et al.* [11] considered reducing the spatial aliasing by making use of a space-varying filter of the light field in an iterative fashion. Li *et al.* [12] considered deblurring directly. Similarly, Chai *et al.* [13] considered the image-based rendering where a minimum sampling rate for light field rendering is presented. Ng *et al.* [6] exterminated the spatial aliasing and considers them as the artifacts introduced by the reconstruction of the light field. As somewhat related, You *et al.*considered utilizing water-drops as lenslets for light field estimation [14].

Light field enables various applications, other than traditional applications in depth estimation [15], [16], Wang *et al.* [17] considered textured material recognition using a 4D light field dataset along with a CNN architecture. However, their method only works for Lambert material where the angular aliasing barely exists. Besides, Li *et al.* [18] considered using saliency detection.

B. Material/BRDF Recognition

Material recognition is an important task for computer vision. Researches in this direction can be roughly divided into two categories: those based on image appearances (e.g., texture), and those based on reflectance property/BRDF. In the first category, Li *et al.* [19] proposed a data-driven method that uses virtual examples to help material recognition. Liu *et al.* [20] proposed a Bayesian framework using low and mid-level features for material recognition. Recently, Sharan *et al.* [21] exploited the perceptually inspired features for this task while Cimpoi *et al.* [22] used deep filter banks for feature extraction. Wang *et al.* [17] proposed a 4D light field dataset and a CNN architecture to recognize materials using 4D data.

Existing methods in the second category exploit the surface reflectance property, i.e., the BRDF property. For accurate BRDF measurement, several works have been done by using professional capture devices [1], [23], and having exact surface BRDFs, material classification can be done nicely [24]. However, devices for BRDF capture are not widely available for

common users. Dana *et al.* proposed a special mirror used for BRDF/BTF measurement [25], and for convenient measurement of spatially varying bidirectional reflectance [26]. By using similar devices, Han *et al.* [27] proposed to measure bidirectional texture reflectance with a kaleidoscope, and Zhang *et al.* [28] captured reflectance disks and proposed a reflectance hashing method for material recognition. Without requiring additional devices, some other methods need to assume varying illuminations [29]–[34], multiview capturing [35], or known object shapes [36]. Overall, recognize surface BRDFs from a single shot is still challenging.

C. Material/BRDF Recognition Dataset

There are various public datasets produced for material recognition. For instance, the Flickr Material Database (FMD) [37] contains 10 categories with 100 images in each category, the Context Database (MINC) contains 3 million real world image patches classified into 23 materials, and the Describable Textures Dataset (DTD) [38] is another popular one. There are also synthetic datasets [39] compose of computer-generated images, and 4D light-field dataset [17] for material recognition. These datasets all focus on images appearances rather than reflectance property.

There are only a few datasets designed for reflectance/BRDF recognition. Zhang *et al.* [28] proposed a reflectance disk dataset which captures intensities of different viewing angles for 20 materials. However, their dataset lacks spatial information and thus is small.

III. Capturing BRDF in 4D Light Field Image

In this section, we aim to formulate the BRDF identification task under a light field setting. First, the basic concept of BRDF is introduced in Section III-A. Then, the light field imaging setup is provided in Section III-B. Finally, the idea of BRDF identification using a single 4D light field image is analyzed.

A. Surface Appearance and BRDF

BRDFs measure the ratio of the reflected radiance from a surface area. It is a function $f(\boldsymbol{\omega}_{in}, \boldsymbol{\omega}_{out})$ of incoming light direction $\boldsymbol{\omega}_{in}$ and outgoing light direction $\boldsymbol{\omega}_{out}$ in a local coordinate system defined by the surface normal. According to its definition, the pixel value captured at a surface point can be computed by

$$I = f(\boldsymbol{\omega}_{in}, \boldsymbol{\omega}_{out})(\boldsymbol{n}^{\mathrm{T}}\boldsymbol{s}), \tag{1}$$

where $\boldsymbol{n}$ is the unit normal vector of the surface point and $\boldsymbol{s}$ indicates the point light source.

As explained above, the BRDF is a 4-dimensional function which can be highly complex. In most cases, we only consider simple isotropic BRDFs, e.g., the commonly used 100 different isotropic BRDFs in the MERL database [1]. Moreover, those isotropic BRDFs can be categorized into several groups, and the BRDFs inside each group appearing similarly. Accordingly, in this paper we propose five typical BRDF types and use them for BRDF identification.

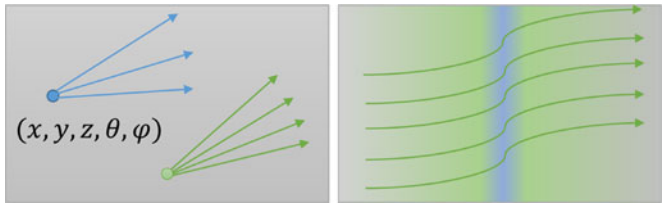


Fig. 2. Light transfer in a media. Left: the power at the given point and given direction can be defined. Right: non-uniform media can cause detours, refraction and reflection of the light.

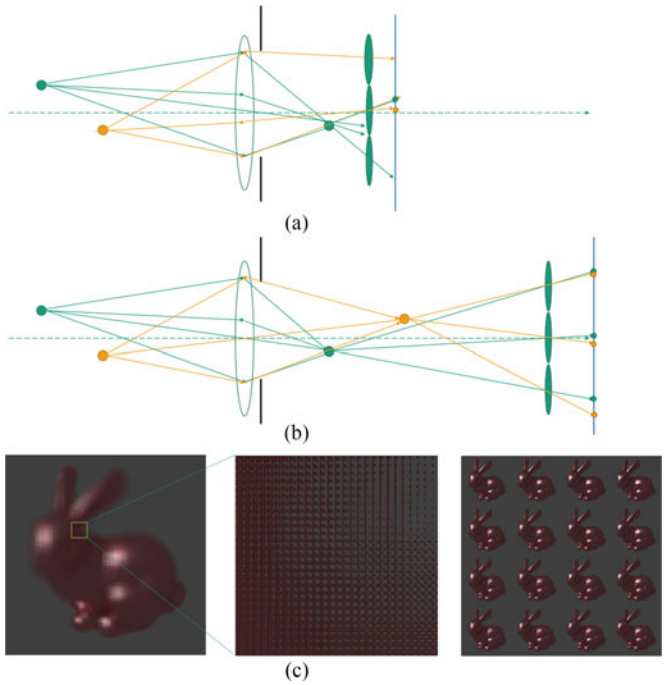


Fig. 3. Light Field Capturing Device (Light Field Cameras) (a) Standard light field camera, the device captures the light field at the 2D place located at the front lens. Light transfers through the front lens, and later the back lens array re-projects the light captured from different directions into the image plane. (b) Equivalence of light field cameras as a set of perspective camera arrays, and one of their captured images. (c) Examples of light field image and multi-view images.

B. Light Field Imaging

We briefly describe the light field imaging technique. This will help understand why light field imaging can benefit the BRDF identification.

Light Field: Light field is a fundamental idea in optics. As illustrated in Fig. 2, light transfers in spaces as a particle. For any media allows a light transfer, it forms a light field. For any given position (x, y, z) in the media, and any given direction (θ, φ), one can observe the field transfer, and in particular, the power (luminescence) of the light transferring at the given point and direction, which is usually defined as $I(x, y, z, \theta, \varphi)$. For a given media with quasi-stable light transferring, the light field is the power of the light transferring among all the positions and directions:

$$I(x, y, z, \theta, \varphi), x, y, z \in \Omega, \tag{2}$$

where ω is the optical media where allows the light transfer.

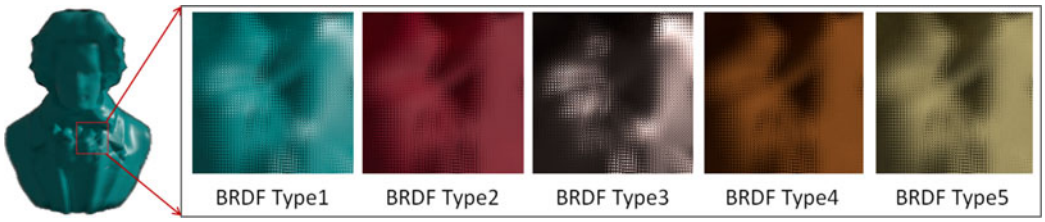


Fig. 4. Different appearances of light field images with different BRDFs. The five BRDF types are: plastic, diffuse, narrow-lobe-metal, wide-lobe-metal, and mid-lobe-metal correspondingly.

Usually, the media is uniform, and light transfers in straight rays. However, in other situations the light transfer undergoes detours, refraction and reflection, as shown in Fig. 2.

Light Field Imaging: As illustrated in Fig. 3, a standard light field camera captures the light field at the front lens, and then re-projects the light from different directions into the image plane via the back lens array. Each lenslet, which can be considered as a perspective camera and captures the light from a certain direction, only covers a portion of all pixels on the sensor due to the limited spatial scale.

As shown in Fig. 3, we can define the light field on the image sensor. As a common practice, we denote the position of lenslets using s, t axis, and the pixel coordinates in each image using u, v axis. Thus we have:

$$I(s,t,u,v), s,t \in \Omega_S, u,v \in \Omega_A, \tag{3}$$

where Ω_S indicates the spatial area, it can be converted between the area on the image sensor and the area on the front lens [40]. Similarly, Ω_A defines the possible angular area, and it is also determined by the distance from the object observed and the front lens aperture.

Assuming the total number of lenslets are S and T, and the number of pixels in each lenslet as U and V. One can formulate the trade-off that $SU = constant$ and $TV = constant$.

Note that the light field imaging is not necessary to be equivalent to the camera array. However, in this paper we use the camera array model to formulate the light field imaging.

C. Light Field of BRDF

Based on the above discussion, we describe the benefit of identifying the BRDFs from light field images.

Referring to (1), the imaging process is determined by the $geometry: \boldsymbol{n}, observingdirection: \boldsymbol{\omega}_{out}$, and the $BRDF: f$ and the $illumination$.

The BRDF identification task can be defined as finding the BRDF which best fits the light field observation.

$$\max_{f \in B} p(I_{LF}|I_f). \tag{4}$$

Where B is the possible BRDF collection, I_f denotes the appearance prior given the BRDF f, and $I_L F$ denotes the observed light field image.

Fig. 4 provides examples of how different BRDFs (plastic, lambertian and metal with different gloss levels) vary the final light field images. Detailed analysis on how BRDF can be effectively captured by the light field image is given below.

Geometry: The light field information enables us to capture the surface geometry from a single image. This is because the light field camera can be considered as perspective camera arrays in some stand [41], [42] to allow shape from light field. Consequently, the surface geometry, especially the surface normal, is an essential factor for measuring BRDFs as described in Section III-A.

Angular Sampling BRDFs are defined as the inbound and outbound light relations, meaning that the angular sampling is crucial. Traditional cameras can only observe one inbound-outbound relation in one image, while a light field image can capture multiple observations in angular domain.

Spatial Resolution: One aspect of light field imaging is that the spatial resolution is not as high as a perceptive camera. However, it is not a substantial problem in BRDF estimation. A light field camera can still have a spatial resolution as high as 1000 by 1000 pixels or more. Which is sufficient for BRDF estimation.

Following the above discussions, in this paper, we propose to tackle the BRDF identification problem by using a light field image. Rather than explicitly measuring the geometric and angular values, we take the advantage of deep convolutional neural networks and learn the most distinguishing features from a single light field image automatically.

IV. Light Field BRDF Dataset

Different from common 2D images, the light field images are neither easy to capture in real world or widely available in Internet. Therefore, for the sake of studying how light field images can benefit the BRDF identification, we provide a novel synthetic light field BRDF dataset in this paper. All images in our dataset are 4D light field images produced by image synthesis. Different combinations of BRDFs, 3D shapes, and illuminations are covered by the dataset.

A. Light Field Image Rendering

We render a large amount 4D light field images by using the following configuration.

BRDFs: We define five typical types of BRDFs, namely, plastic, diffuse, narrow-lobe-metal, wide-lobe-metal, and mid-lobe-metal BRDFs, as seen in the top-left of Fig. 5. Each BRDF type also contains three representative BRDFs, which are similar but not identical. This is to ensure the diversity inside each type. As a result, there are totally fifteen different BRDFs in our dataset.

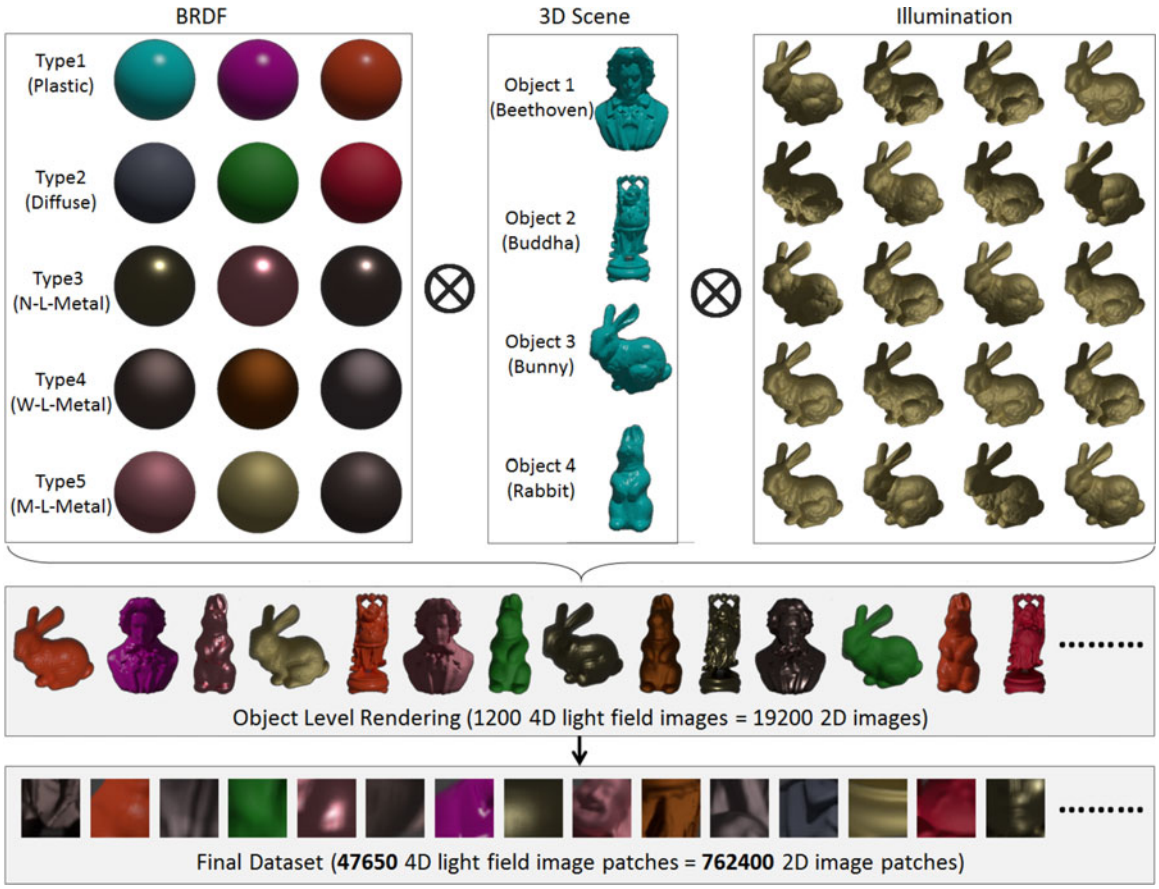


Fig. 5. The proposed dataset for the BRDF identification task. The dataset is generated via image synthesis with five typical types of BRDFs, namely, plastic, diffuse, narrow-lobe-metal, wide-lobe-metal, and mid-lobe-metal BRDFs, four typical 3D objects, namely, Beethoven, Buddha, Bunny and Rabbit, and twenty different illumination conditions. Totally 1200 4D light field images are synthesized which comprise 19200 2D images. Finally, we carefully crop out 47650 4D light field image patches (comprising 762400 2D image patches) from the images to form our dataset.

3D objects: For rendering the images, we use four different and complex 3D objects, including beethoven, happy buddha, bunny, and rabbit. They are easily accessible from the Internet, as visualized in Fig. 5.

Illumination: We design twenty different illumination conditions. Each illumination condition is produced by a distant point light source with a constant intensity and also an ambient light which is uniform from all directions. The directions of the point light sources roughly cover the hemisphere in the camera side. It is important to ensure the diversity of light source directions since different light source directions could produce significantly different surface appearances. The illuminations are also visualized in Fig. 5.

Light Field Imaging Setting: We render each of the 4D light field images by assuming each lenslet covers 4×4 pixels on the image sensor, and thus the angular resolution of the light field image is 4×4. By setting the resolution of any image captured by a lenslet to be 512×512, the final resolution of the 4D light field image is 2048×2048. Note that all the 16 lenslets/cameras have the same distance to the 3D model, and the angular difference of viewing directions of adjacent lenslets/cameras is set to be 2°.

Overall, by considering all the combinations of BRDFs, 3D objects, and illuminations, we render $15 \times 4 \times 20 = 1200$ different 4D light field images, and use them to constitute our dataset. If considering the 2D sub-images captured by different lenslets, our dataset contains $1200 \times 16 = 19200$ 2D images. Some representative images are shown in Fig. 5, whose appearances differ from each other due to different surface BRDFs, 3D shapes, and the illuminations.

B. Final Dataset With Light Field Image Patches

In order to use the rendered light field data in a more efficient way, we extract image patches from the entire images.

First, each 4D light field image is decomposed into 16 2D images. Then, images patches with a size of 64×64 are extracted at the same position of the 16 2D images. Finally, by merging all the 16 2D image patches, the 4D light field image patch with a size of 256×256 is produced.

The positions where the 64×64 image patches are extracted from the 2D image are determined as follows. For each 2D image, we traverse through its different pixel positions by a step of 4 pixels. At each valid pixel position, a crop should be done if more than 95% area of the cropped image patch can be occupied by the object/foreground region. Because difference in BRDF/illumination does not affect the foreground/background distribution in the image, the numbers of cropped image patches must be the same for images rendered for the same 3D object. In particular, Table I shows the numbers of image

TABLE I
NUMBERS OF IMAGE PATCHES THAT CAN BE CROPPED OUT FROM A SINGLE IMAGE OF DIFFERENT 3D OBJECTS

3D object	Beethoven	Buddha	Bunny	Rabbit	Total
# of pathes per image	3062	1458	3366	1644	9530

patches that can be cropped out from one image of different 3D objects.

Theoretically, we can produce $9530 \times 15 \times 20 = 2859000$ image patches as described above. However, notice that there is a change that two patches share the same 3D shape (the same crop position) and the same (or similar) illumination, while their BRDFs are not the same but belonging to the same BRDF type. In such cases, the appearances of the two patches become quite similar. In order to avoid such cases, we propose the following strategy. At each crop position, we randomly select five BRDFs with different BRDF types; at the same time, we also randomly select five illumination conditions. In other words, we only produce five image patches (out of all the 300 possible ones) at each crop position, and their appearances are expected to be quite different. In this way, $9530 \times 5 = 47650$ 4D image patches are obtained, which are ensured to be significantly different from each other.

Each 4D light field image patch is then annotated by its BRDF type label. They constitute our final dataset. To our knowledge, this is the first light field BRDF dataset of this large for the BRDF identification task.

V. CNN ARCHITECTURE FOR BRDF IDENTIFICATION

Following the theoretical analysis in Section III, we propose two CNN architectures for BRDF identification in this section.

A. System Overview

The input to the CNNs is a 4D light field image. In particular, we design two different schemes to process the input in the early stage of the CNN.

StackNet: We convert the light field image into a 3D image volume before sending it to the CNN. In particular, the original light field image is re-organized such that the first two dimensions represent the spatial observation s, t and the angular observation u, v is mapped to the third dimension. Intuitively, the 3D image volume is a stack of 16 gray-scale images of size 64×64 , each of which can be considered as being captured by one of the 16 lenslets, as shown in Fig. 6. Then, the 3D volume is input to a convolutional layer with 64 kernels. Finally, the output is sent to the rest of the network, which is described later.

Ang-convNet: To better utilize the angular information, we propose an Ang-convNet that processes the 4D light field image as a whole. In this method, we introduce 64 so-called angular convolutional kernels (ACKs). Each ACK is a 4 by 4 filter. It differs from common convolutional kernels in that it only filters the angular area in the (gray-scale) light field image. In other words, the stride of convolution operations is 4 in both the horizontal and vertical dimensions. Therefore, the filtering result is a $64 \times 64 \times 64$ data volume, as that in the StackNet. The output is then sent to the rest of the network, as shown in Fig. 6. This method directly learns to handle angular information in the light field image, and thus is expected to have better performance.

As mentioned above, both the StackNet and the Ang-convNet share a large portion of the same network structure, as shown in Fig. 6. In particular, the shared part contains two convolution layers, each of which is followed by a ReLu layer, a normalization layer and a max pooling layer. These operations convert the $64 \times 64 \times 64$ input data into a $16 \times 16 \times 128$ feature, which is then reshaped to form a 1D feature vector. Then, the feature vectors dimension is further reduced to 384×1, 192×1, and 5×1, respectively, after two fully connected layers and a SoftMax classifier. By checking the final output, the BRDF type label can be determined. Note that the stride of all convolution operations in the network is set to 1, and the padding type is zero padding. The image size is changed only during the max pooling operations.

B. Training the CNNs

After finishing the data collection and the CNN structure design, we train the networks by using the ground truth data. In particular, we use the softmax cross-entropy loss function defined as

$$J = -\sum_{i=1}^{n} \sum_{c \in C} t_{ic} \cdot log(y_{ic}), \tag{5}$$

where i indicates different samples and c indicates different BRDF type labels. Note that t_{ic} is 1 only when sample i belongs to BRDF type c, and the softmax output y_{ic} is the probability that sample i belongs to BRDF type c. The loss function computes the difference between the output label and the ground truth, which is iteratively minimized by using the gradient descent algorithm. Besides, we choose to add a dropout layer after each fully connected layer to discard a portion of learnt weights randomly. This helps avoid local optimal, according to Krizhevsky *et al.* [43]. Note that during test, the dropout layers are removed. By setting the learning rate to 0.001, the network learns the relation between the input light field image and its corresponding BRDF type label. The network is then used for the BRDF identification task.

VI. EXPERIMENTS

In this section, we evaluate the proposed CNN-based BRDF identification methods on our light field BRDF dataset. The dataset contains 47650 4D light field image patches, which randomly cover 5 BRDF categories, 9530 different shapes, and 20 illuminations. Our proposed strategy in Section IV-B ensures that any two patches rarely get a chance to share a similar appearance, and thus those 47650 patches are unique.

For experiments, we randomly choose 40000 4D light field image patches as training samples, and the rest 7650 patches as test samples. The training samples are then sent to the StackNet, Ang-convNet and two baseline methods for training, and the test samples are used for evaluation. BRDF type identification accuracy is finally computed from the test output.

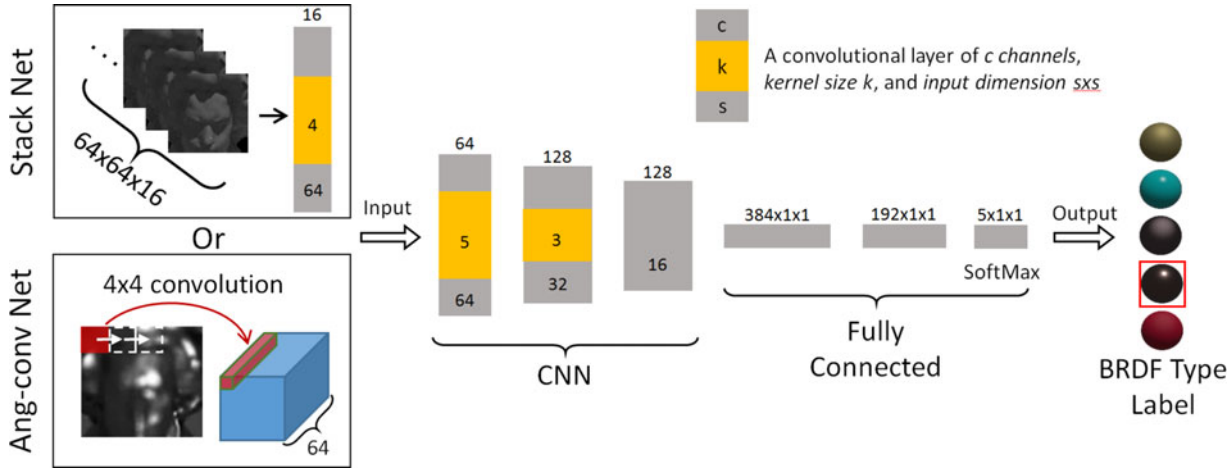


Fig. 6. The proposed CNN structures to utilize 4D light field input for BRDF identification. The StackNet first decomposes one light field image into 16 2D images and then put them into a convolutional layer, while the Ang-convNet uses 4 by 4 angular filters to encode the angular information in the 4D light field image. Then, the output of either the StackNet or the Ang-convNet is passed to the following CNN with 2 typical convolutional layers and 2 typical fully connected layers for BRDF type identification.

TABLE II
ACCURACY W.R.T. BRDF TYPES

	BRDF 1	BRDF 2	BRDF 3	BRDF 4	BRDF 5	Total
Random Forests	56.3%	**75.1%**	90.2%	77.6%	57.9%	71.5%
SingleNet	69.8%	65.1%	82.6%	68.9%	69.8%	61.6%
StackNet	79.9%	74.1%	90.9%	62.0%	70.7%	75.4%
Ang-convNet	**87.6%**	74.7%	**94.9%**	**82.7%**	**78.4%**	**83.7%**

TABLE III
ACCURACY W.R.T. 3D OBJECTS (WHERE THE INPUT PATCH COMES FROM)

	Object 1	Object 2	Object 3	Object 4	Total
Random Forests	68.8%	68.6%	72.5%	77.1%	71.5%
SingleNet	70.7%	70.5%	66.8%	74.1%	69.8%
StackNet	71.6%	76.9%	75.8%	80.8%	75.4%
Ang-convNet	**82.8%**	**84.1%**	**84.4%**	**83.3%**	**83.7%**

Note that our experiments are conducted by using the proposed synthetic dataset. This is because the synthetic dataset can provide different combinations of shapes, BRDFs, and illuminations. Besides, the BRDFs from the MERL dataset were accurately measured from real surfaces, which ensures that the synthetic data can approximate the real world data with high precision. Therefore, our results obtained in this section can also reflect the performance of the proposed methods on the real world data, if a proper light field camera can be used.

A. Baseline Methods

We conduct numerical evaluations for the proposed StackNet and Ang-convNet, and also compare their accuracy with those of two baseline methods. The baseline methods include:

SingleNet: We propose the SingleNet as a baseline method for comparison. The SingleNet differs from the proposed StackNet and Ang-convNet in that it only accepts a single 2D image/image patch as input, rather than a 4D light field image/image patch. Therefore, it can provide evidence of whether light field imaging can benefit the BRDF identification task. In particular, the SingleNet uses the same CNN architecture as shown in Fig. 6 after removing the procedures in the two boxes that handle light field images in the StackNet and Ang-convNet. Besides, the first CNN layer has to be modified to accept a single image as input.

Random Forests (RF): Random forests is an effective ensemble learning method for classification, regression and other tasks.

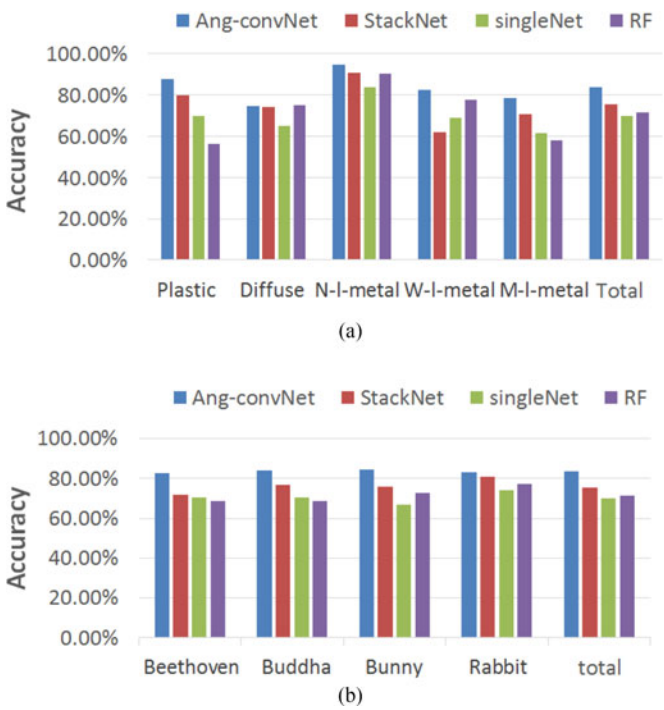
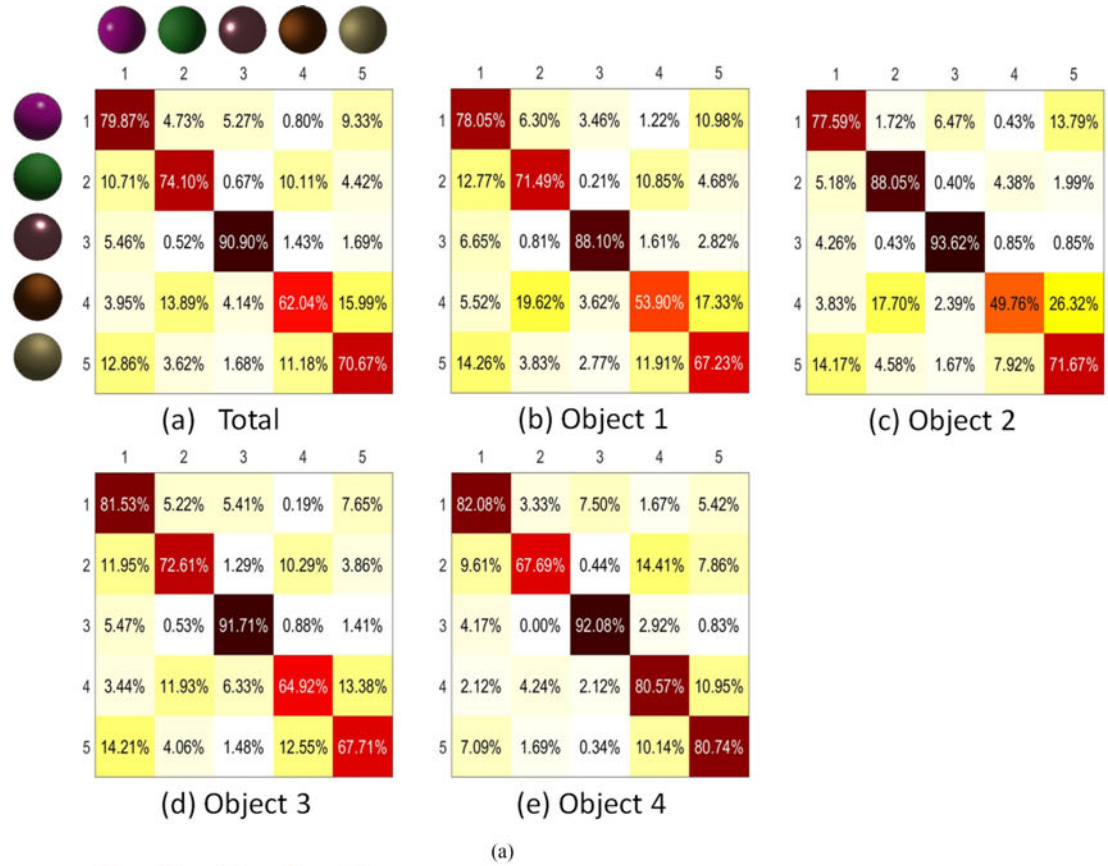
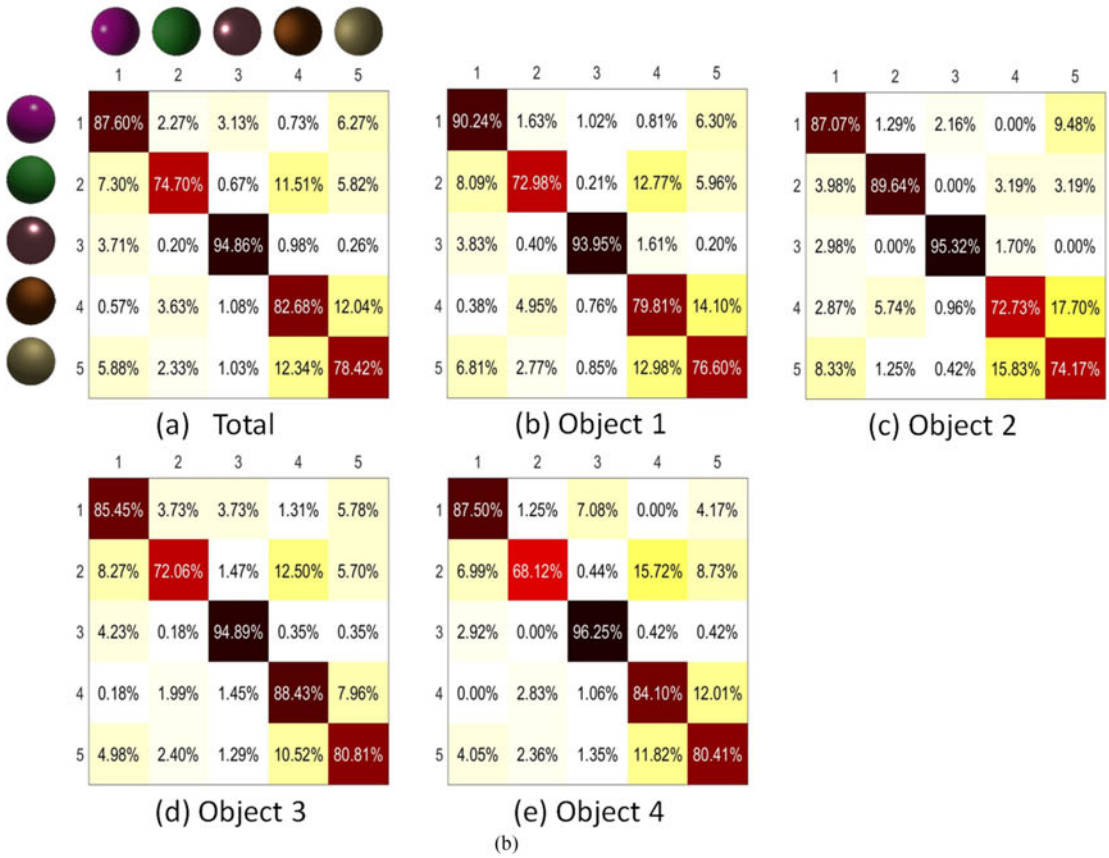


Fig. 7. Comparison of identification accuracy between Ang-ConvNet and StackNet. (a) Accuracy w.r.t. different BRDF types. (b) Accuracy w.r.t. different 3D objects. Five BRDF types include plastic, diffuse, narrow-lobe-metal, wide-lobe-metal, and mid-lobe-metal respectively.



(a)



(b)

Fig. 8. Confusion matrices of BRDF identification results. Top: StackNet. Bottom: Ang-convNet. Five BRDF types are plastic, diffuse, narrow-lobe-metal, wide-lobe-metal, and mid-lobe-metal respectively. (a) Confusion matrix of the StackNet's results. (b) Confusion matrix of the Ang-convNet's results.

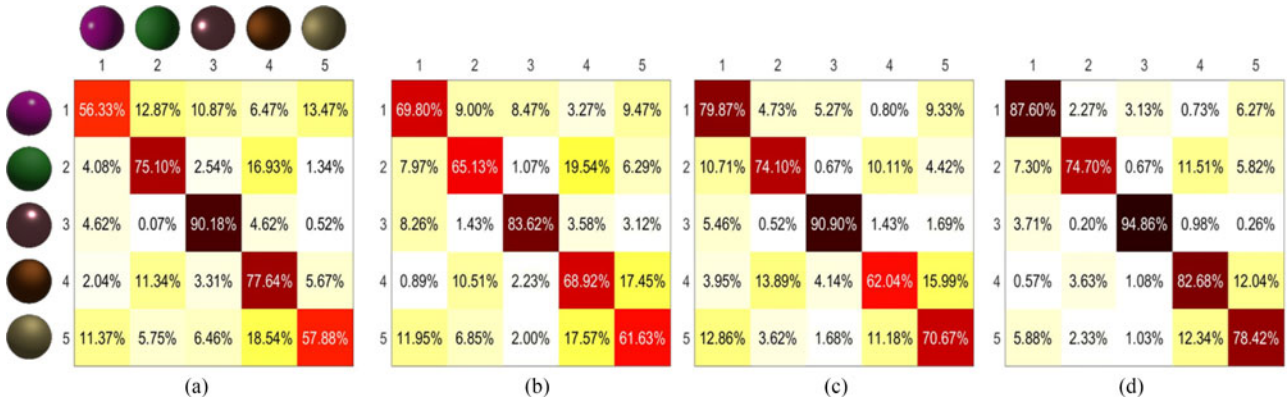


Fig. 9. Comparison between different methods on BRDF type identification accuracy with respect to each BRDF type.

It operates by constructing decision trees at training stage and outputting the class label as the classification result. Random forests can avoid disadvantages of using common decision trees such as the overfitting problem. It is a representative method for solving problems like ours and has been adopted by related researches for thousands of times.

B. Comparison of Different Methods

We conduct experiments on the proposed dataset by using the proposed methods, i.e., StackNet and Ang-convNet, and the two baseline methods, i.e., SingleNet and the Random Forests. The results are shown and compared in Table II. Overally speaking, the StackNet and Ang-convNet significantly outperform the baseline methods and show good performance in the task of BRDF identification (75.4% for StackNet and 83.7% for Ang-convNet). This proves the advantage of using light field information for BRDF identification. While examining their performance for each individual BRDF type, the Ang-convNet outperforms the StackNet especially for BRDF type 4. This reflects the fact that treating the light field image as a whole and applying the proposed angular filtering can help exploit the angular information, which is important for the identification of certain types of BRDF.

We also evaluate the performance of different methods w.r.t. different 3D objects where the test image patches come from. Results are provided in Table III for all of the four methods, from which we can conclude that 1) the proposed methods constantly outperform the baseline methods. 2) different 3D objects do affect the BRDF type identification accuracy for the Random Forests, the SingleNet and the StackNet methods. In particular, Object 4 seems to be the easiest one partially because it has the smoothest surface. 3) On the other hand, the Ang-convNet not only achieves the highest average accuracy, but also performs very stably for all objects. This demonstrates its ability to capture the essential feature for BRDF type identification without being affected by other factors like different shapes.

For better visualization of the results, we plot the accuracy of all the four methods in Fig. 7 and make a comparison. The proposed methods outperform the baseline methods, and the Ang-convNet achieves the highest accuracy and the number is quite satisfactory.

C. Detailed Analysis on BRDF Type Identification Accuracy

This section provides detailed discussions on the BRDF type identification results for Ang-ConvNet and StackNet, as visualized in the confusion matrices in Fig. 8.

StackNet: Several observations can be made from the results of the StackNet. 1) In general, BRDF type 3 achieves the highest identification accuracy of 90.9%, while the number for BRDF type 4 is only 62.0%; 2) More than 15% samples of BRDF type 4 are wrongly classified to be BRDF type 5. This is because BRDF type 4 and 5 has the highest similarity with each other, as shown in Fig. 8; 3) If considering different Objects, it can be seen that Object 4, which has smoothest surface among all, is an "easy one" for identifying BRDF type 4, but a "difficult one" for BRDF 2. This reveals the fact that the performance of the StackNet can be affected by the 3D shapes, in a BRDF type dependent manner.

Ang-convNet: Generally speaking, the Ang-convNet achieves significantly better results than the StackNet in nearly all the cases. However, similar to the StackNet, different objects still affect the BRDF type identification performance. Another observation is that, BRDF type 2, rather than BRDF type 4 in the StackNet case, gets the highest chance among all BRDF types to be mis-identified. In other words, replacing the StackNet by the Ang-convNetBRDF results in the least improvement for BRDF type 2.

Overall, as shown in Fig. 9, the Ang-convNet shows better capability in BRDF type identification. It proves that the angular filtering process applied to the 4D light field images does capture the essential information about the BRDF type.

D. Failure Case Study

In this section, we show some individual cases where the BRDF identification outputs wrongly in our experiment. As shown in Fig. 10, we pick up four individual cases for illustration. In particular, the Ang-convNet is most likely to misidentify the 1) BRDF type 5 to Type 4, 2) BRDF type 4 to Type 5 and 3) BRDF type 2 to type 4, with error rates of 12.34%, 12.04% and 11.51%, respectively. As visualized in Fig. 5, the BRDF type 4 and 5 are hard to distinguish even for human beings, and they also appear similar with BRDF type 2 in many



Fig. 10. Failure cases in BRDF type identification. The first three rows show the failure cases with the highest possibilities, while the bottom row shows the most unlikely case. Note that the wrong outputs are additionally synthesized and they do not exist in the dataset. Also note that when comparing different BRDF types, the difference in color is ignored.

cases. Therefore, the corresponding images shown in Fig. 10 look similar and this explains the high error rates.

On the other hand, the chance that BRDF Type 3 is misidentified to Type 2 is only 0.20%, which is the lowest in our experiments. This is confirmed by the obvious difference in appearance of the two BRDF types. Only several failure cases of this kind can be found as shown in the bottom row of Fig. 10, where the image patch of BRDF type 3 happens to have no specular region and this makes it similar to BRDF type 2. This reveals the difficulty of BRDF type identification from only a single image.

VII. Conclusion

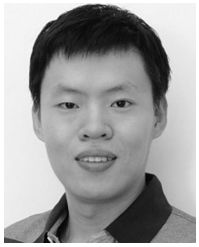
In this paper, we propose novel methods for BRDF identification by using a single light field image. The key idea is that a light field image captures both spatial and angular information through a single shot, and the additional angular information enables effective samplings of the 4D BRDF. Following the idea, a novel light field BRDF dataset and two CNN-based architectures for BRDF identification from a single 4D light field image are provided. Specifically, to cope with the 4D light field image, a StackNet and an Ang-convNet are designed respectively. The StackNet re-organizes the original light field image into a 3D image volume according to its angular observation, and the Ang-convNet uses a 4 by 4 filter to encode the angular information. The dataset is generated by using different 3D models, BRDFs and illuminations. Totally 47650 high quality 4D light field image patched are produced, making the dataset bigger than existing light field datasets. Systematical experiments demonstrate that the proposed methods outperform conventional methods.

Future works include the following. First, besides the synthetic dataset, one can also collect real images by using proper light field cameras for further investigation. Second, while the proposed methods work in a classification manner, they can also be modified for BRDF parameter regression and benefit related applications [44], [45].

References

[1] W. Matusik, H. Pfister, M. Brand, and L. McMillan, "A data-driven reflectance model," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 759–769, 2003.
[2] G. Wu *et al.*, "Light field image processing: An overview," *J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.
[3] T. Georgiev and A. Lumsdaine, "Depth of field in plenoptic cameras," in *Proc. Eurographics*, 2009, pp. 5–8.
[4] E. H. Adelson and J. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 99–106, Feb. 1992.
[5] M. Levoy and N. Hanrahan, "Light field rendering," in *Proc. SIGGRAPH*, 1996, pp. 31–42.
[6] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Comput. Sci. Tech. Rep.*, vol. 2, no. 11, pp. 1–11, 2005.
[7] T. Georgiev, Z. Yu, A. Lumsdaine, and S. Goma, "Lytro camera technology: Theory, algorithms, performance analysis," *Proc. SPIE*, vol. 8667, 2013, Art. no. 86671J.
[8] G. Lippmann, "Épreuves réversibles donnant la sensation du relief," *J. Phys. Theory Appl.*, vol. 7, no. 1, pp. 821–825, 1908.
[9] H. E. Ives, "A camera for making parallax panoramagrams," *J. Opt. Soc. Amer.*, vol. 17, no. 6, pp. 435–439, 1928.
[10] J. Stewart, J. Yu, S. J. Gortler, and L. McMillan, "A new reconstruction filter for undersampled light fields," in *Proc. 14th Eurograph. Workshop Rendering*, 2003, pp. 150–156.
[11] T. E. Bishop and P. Favaro, "Full-resolution depth map estimation from an aliased plenoptic light field," in *Proc. 10th Asian Conf. Comput. Vis.*,, 2011, pp. 186–200.
[12] F. Li, J. Yu, and J. Chai, "A hybrid camera for motion deblurring and depth map super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[13] J. X. Chai, X. Tong, S. C. Chan, and H. Y. Shum, "Plenoptic sampling," in *Proc. SIGGRAPH*, 2000, pp. 307–318.
[14] S. You, R. T. Tan, R. Kawakami, Y. Mukaigawa, and K. Ikeuchi, "Waterdrop stereo," 2016, pp. 1–12, arXiv:1604.00730.
[15] E. Ekmekcioglu, V. Velisavljevic, and S. T. Worrall, "Content adaptive enhancement of multi-view depth maps for free viewpoint video," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 2, pp. 352–361, Apr. 2011.
[16] K. Li, Q. Dai, W. Xu, and J. Yang, "Temporal-dense dynamic 3-d reconstruction with low frame rate cameras," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 5, pp. 447–459, Sep. 2012.
[17] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4D light-field dataset and CNN architectures for material recognition," in *Proc. Eur. Conf. Comput. Vis.*, New York, NY, USA: Springer, 2016, pp. 121–138.
[18] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2806–2813.
[19] W. Li and M. Fritz, "Recognizing materials from virtual examples," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 345–358.
[20] C. Liu, L. Sharan, E. H. Adelson, and R. Rosenholtz, "Exploring features in a Bayesian framework for material recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 239–246.

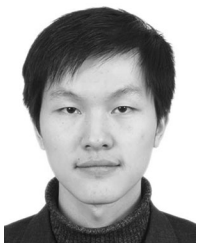
[21] L. Sharan, C. Liu, R. Rosenholtz, and E. Adelson, "Recognizing materials using perceptually inspired features," *Int. J. Comput. Vis.*, vol. 103, no. 3, pp. 348–371, 2013.
[22] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3828–3836.
[23] M. K. Johnson and E. H. Adelson, "Retrographic sensing for the measurement of surface texture and shape," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1070–1077.
[24] O. Wang, P. Gunawardane, S. Scher, and J. Davis, "Material classification using BRDF slices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2805–2811.
[25] K. J. Dana, "BRDF/BTF measurement device," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2001, vol. 2, pp. 460–466.
[26] K. Dana and J. Wang, "Device for convenient measurement of spatially varying bidirectional reflectance," *J. Opt. Soc. Amer. A*, vol. 21, pp. 1–12, 2004.
[27] J. Y. Han and K. Perlin, "Measuring bidirectional texture reflectance with a kaleidoscope," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 741–748, 2003.
[28] H. Zhang, K. Dana, and K. Nishino, "Reflectance hashing for material recognition," in *Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3071–3080.
[29] C. Liu and J. Gu, "Discriminative illumination: Per-pixel classification of raw materials based on optimal projections of spectral BRDF," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 86–98, Jan. 2014.
[30] A. Hertzmann and S. Seitz, "Example-based photometric stereo: Shape reconstruction with general, varying BRDFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1254–1264, Aug. 2005.
[31] F. Lu, I. Sato, and Y. Sato, "Uncalibrated photometric stereo based on elevation angle recovery from BRDF symmetry of isotropic materials," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 168–176.
[32] F. Lu, Y. Matsushita, I. Sato, T. Okabe, and Y. Sato, "From intensity profile to surface normal: Photometric stereo for unknown light sources and isotropic reflectances," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 1999–2012, Oct. 2015.
[33] S. Li and B. Shi, "Photometric stereo for general isotropic reflectances by spherical linear interpolation," *Opt. Eng.*, vol. 54, no. 8, pp. 83–104, 2015.
[34] F. Lu, X. Chen, I. Sato, and Y. Sato, "Symps: BRDF symmetry guided photometric stereo for shape and light source estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. 99, pp. 1–14, 2017. doi: 10.1109/TPAMI.2017.2655525.
[35] G. Oxholm and K. Nishino, "Multiview shape and reflectance from natural illumination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2163–2170.
[36] S. Lombardi and K. Nishino, "Single image multimaterial estimation," in *Proc. 2012 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 238–245.
[37] L. Sharan, R. Rosenholtz, and E. Adelson, "Material perception: What can you see in a brief glance?" *J. Vis.*, vol. 9, no. 8, pp. 784–784, 2009.
[38] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3606–3613.
[39] M. Weinmann, J. Gall, and R. Klein, "Material classification based on training data synthesized using a BTF database," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 156–171.
[40] Y. Liu, Q. Dai, and W. Xu, "A real time interactive dynamic light field transmission system," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2006, pp. 2173–2176.
[41] C. Wu, Y. Liu, Q. Dai, and B. Wilburn, "Fusing multiview and photometric stereo for 3d reconstruction under uncalibrated illumination," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 8, pp. 1082–1095, Aug. 2011.
[42] C. Wu, K. Varanasi, Y. Liu, H.-P. Seidel, and C. Theobalt, "Shading-based dynamic shape refinement from multi-view video under general illumination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1108–1115.
[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. 25*, 2012, pp. 1097–1105.
[44] B. Shi, Z. Wu, Z. Mo, D. Duan, S.-K. Yeung, and P. Tan, "A benchmark dataset and evaluation for non-Lambertian and uncalibrated photometric stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3707–3716.
[45] X. Chen, H. Wu, X. Jin, and Q. Zhao, "Face illumination manipulation using a single reference image by adaptive layer decomposition," *IEEE Trans. Image Process.*, vol. 22, no. 11, pp. 4249–4259, Nov. 2013.



Feng Lu received the B.S. and M.S. degrees in automation from Tsinghua University, Beijing, China, in 2007 and 2010, respectively, and the Ph.D. degree in information science and technology from the University of Tokyo, Tokyo, Japan, in 2013. He is currently a Professor in the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include 3-D shape recovery, reflectance analysis, and human gaze analysis.



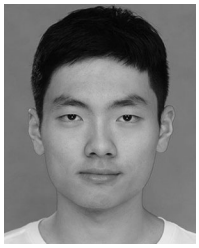
Lei He received the B.S. degree in computer science from Jilin University, Changchun, China, in 2012. He is currently working toward the M.S. degree in computer science at Beihang University, Beijing, China. His research interests include photometric methods in computer vision and 3-D modeling.



Shaodi You received the Ph.D. and M.E. degrees from the University of Tokyo, Tokyo, Japan, in 2015 and 2012, respectively, and the Bachelor's degree from Tsinghua University, Beijing, China, in 2009. He is currently a Research Scientist in Data61, (Previously known as NICTA) Canberra Research Lab and an Adjunct Lecturer in the College of Engineering and Computer Science, Australian National University, Canberra, Australia. His research interests include physics based vision, perception based vision and learning, and 3-D geometry.



Xiaowu Chen received the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2001. He is currently a Professor in the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China. His current research interests include virtual reality, computer graphics, and computer vision.



Zhixiang Hao received the B.S. degree in computer science from Xidian University, Xi'an, China, in 2016. He is currently working toward the Master's degree in the School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests include computer vision and computer graphics.

Super Resolution of Light Field Images Using Linear Subspace Projection of Patch-Volumes

Reuben A. Farrugia, *Member, IEEE*, Christian Galea, *Student Member, IEEE*, and Christine Guillemot, *Fellow, IEEE*

***Abstract*—Light field imaging has emerged as a very promising technology in the field of computational photography. Cameras are becoming commercially available for capturing real-world light fields. However, capturing high spatial resolution light fields remains technologically challenging, and the images rendered from real light fields have today a significantly lower spatial resolution compared to traditional two-dimensional (2-D) cameras. This paper describes an example-based super-resolution algorithm for light fields, which allows the increase of the spatial resolution of the different views in a consistent manner across all subaperture images of the light field. The algorithm learns linear projections between subspaces of reduced dimension in which reside patch-volumes extracted from the light field. The method is extended to cope with angular super-resolution, where 2-D patches of intermediate subaperture images are approximated from neighboring subaperture images using multivariate ridge regression. Experimental results show significant quality improvement when compared to state-of-the-art single-image super-resolution methods applied on each view separately, as well as when compared to a recent light field super-resolution techniques based on deep learning.**

***Index Terms*—Dimensionality reduction, learning, light-fields, super-resolution.**

I. Introduction

LIGHT fields are densely sampled high-dimensional signals containing information about the light rays interacting with the scene. Compared to classical 2D images, light fields capture the intensity values along each ray and not only the sum of intensities of rays reaching each image point. This rich description of a scene enables advanced creation of images with a number of functionalities such as refocusing, extended depth of field, different view point rendering and depth reconstruction which can be very useful for image/video post-production editing and cinematic virtual reality applications [1], [2]. Many light fields acquisition devices have been recently designed, going from arrays of cameras capturing the scene from slightly different viewpoints [2], to single cameras mounted on moving gantries and plenoptic cameras.

Plenoptic cameras, now commercially available, use an array of micro-lenses placed between the sensor and the main lens, to acquire both spatial and angular information of light rays from a single capture [1]. However, since the sensor resolution is limited, it is difficult to have both a dense angular and spatial light field sampling. The angular sampling is related to the number of sensor pixels located behind each microlens, while the spatial sampling is related to the number of microlenses. This trade-off between angular and spatial resolution leads to a significantly lower spatial resolution compared to traditional 2D cameras [1].

Research effort has already been dedicated to this angular and spatial resolution trade-off in light field imaging. For a comprehensive overview of light field image processing techniques, including super-resolution, please refer to [3]. A first category of approaches consists in super-resolving the rendered images, *i.e*. enhancing spatial super-resolution, as in [4] where the authors simultaneously estimate a super-resolved depth map and the all-in-focus image. In the same vein, a light field reconstruction approach is proposed in [5], where the de-multiplexed sub-aperture images are first interpolated with barycentric interpolation to adapt to the hexagonal layout of the micro-lenses, and then refined using pixels of neighboring views using ray interpolation. The rendered image is further enhanced using the dictionary-based super-resolution method proposed in [6] for 2D images.

A second category of methods directly super-resolve the light field to enhance both the spatial and angular resolution. Based on an image formation model of the plenoptic camera, the authors in [7] cast the high-resolution (HR) light field estimation in a Bayesian framework assuming a depth-dependent blurring kernel. A method based on a variational framework is proposed in [8] to synthesize super-resolved novel views with the help of disparity information estimated on the epipolar plane images. A patch-based technique is proposed in [9] to enhance the spatial resolution of light fields. The high resolution 4D-patches are estimated using a linear minimum mean square error (LMMSE) estimator, assuming a disparity-dependent Gaussian Mixture Model (GMM) for the patch structure. More recently, the authors in [10] adopt a deep convolutional neural network (DCNN) to perform both spatial and angular super-resolution. This method first employs a spatial DCNN similar to [11] followed by an angular DCNN which synthesizes intermediate views. A cascade of two DCNNs is used in [12] which exploits disparity information to synthesize intermediate sub-aperture

Manuscript received November 30, 2016; revised April 12, 2017 and July 18, 2017; accepted August 15, 2017. Date of publication August 30, 2017; date of current version October 23, 2017. This was supported in part by the EU H2020 Research and Innovation Programme under Grant 694122 (ERC advanced grant CLIM). The guest editor coordinating the review of this paper and approving it for publication was Prof. Yebin Liu. *(Corresponding author: Reuben A. Farrugia.)*

R. A. Farrugia and C. Galea are with the Department of Communications and Computer Engineering, University of Malta, Msida MSD2080, Malta (e-mail: reuben.farrugia@um.edu.mt; christian.galea.09@um.edu.mt).

C. Guillemot is with the Institut National de Recherche en Informatique et en Automatique, Rennes 35042, France (e-mail: christine.guillemot@intria.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2747127

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

images. A sparse representation approach that can be extended for light-field super-resolution was presented in [13].

In this paper, we propose an example-based spatial super-resolution technique, which, to maintain consistency across all sub-aperture images of the light field, operates on 3D stacks (called *patch-volumes*) of 2D-patches, extracted from the different sub-aperture images. The patches forming the 3D stack are either co-located patches or best matches across sub-aperture images. A dictionary of *examples* is first constructed by extracting, from a training set of high- and low- resolution light fields, pairs of high- and low-resolution *patch-volumes*. These *patch-volumes* are of very high dimension ($q \times q \times n$), where n is the number of sub-aperture images and $q \times q$ is the size of each 2D-patch. Nevertheless, they contain a lot of redundant information, hence actually lie on subspaces of lower dimension. The low- and high-resolution *patch-volumes* of each pair can therefore be projected on their respective low- and high-resolution subspaces. In the sequel, the projection is performed using Principal Component Analysis (PCA). The dictionary of pairs of projected *patch-volumes* (the *examples*) map locally the relation between the high-resolution *patch-volumes* and their low-resolution (LR) counterparts. A linear mapping function is then learned, using Multivariate Ridge Regression (RR), between the subspaces of the low- and high-resolution *patch-volumes*. Each overlapping *patch-volume* of the low-resolution light field can then be super-resolved by a straight application of the learned mapping function.

The above method, which will be referred to as PCA + RR in the sequel, assumes that the 2D collocated patches extracted from all sub-aperture images to form a given *patch-volume*, are well aligned. This may not be the case when large disparities exist across sub-aperture images, depending on the depth of the scene and of the capturing device. For light fields exhibiting large disparities, the above method is further improved by using block matching (BM) to form the *patch-volumes* with the best-matching patches across all sub-aperture images, instead of simply taking collocated patches. An iterative procedure is proposed where a different sub-aperture image is chosen as anchor at each iteration to form the *patch-volumes*. This method will be referred to as BM + PCA + RR.

By exploiting the light field structure, where sub-aperture images represent different viewpoints from the same scene, it becomes evident that a global interpolation function to cater for all views will be sub-optimal. The DCNN scheme presented in [10] models the angular super-resolution function based on the available neighbouring sub-aperture images independently from the viewpoint being approximated. Instead, we model the angular super-resolution problem using multi-linear models, *i.e.* a linear model for every missing sub-aperture image to be synthesized. This is done by exploiting the local neighbouring sub-aperture image and also the angular location of the sub-aperture image being approximated.

Experimental results with both synthetic light fields (from the HCI dataset) and real light fields (from the Stanford and the INRIA datasets) show that the proposed method outperforms state-of-the-art single-image super-resolution methods [14]–[16] applied on each sub-aperture image separately, and significantly outperfoms prior light field super-resolution algorithms for both spatial [9], [10] and angular [10], [12] super-resolution. Average PSNR gains of 3.7 dB, 4.2 dB and 2.1 dB on the HCI, Stanford and INRIA datasets respectively were obtained for spatial super-resolution using BM + PCA + RR compared to the recent DCNN scheme [10] which obtained the second best results in our experiments. Moreover, subjective results clearly show that the proposed angular super-resolution method provides sharper synthesized intermediate sub-aperture images when compared to those obtained using the DCNN based schemes [10], [12].

The rest of the paper is organized as follows. After introducing the notations in Section II, we describe the different steps of the proposed algorithm in Section III. Section IV discusses the simulation results with different types of light fields and provide the final concluding remarks in Section VI.

II. Light Fields Notation

We consider here the simplified 4D representation of light fields called 4D light field in [17] and lumigraph in [18], describing the radiance along rays by a function $L(x, y, s, t)$ where the pairs (x, y) and (s, t) respectively represent spatial and angular coordinates respectively. The light field can be seen as capturing an array of viewpoints (called sub-aperture images) of the scene with varying angular coordinates (s, t). The different views will be denoted here by $\mathbf{V}_{s,t} \in \mathbb{R}^{X,Y}$, where X and Y represent the vertical and horizontal dimension of each sub-aperture image. Each sub-aperture image corresponds to a fixed pair of (s, t).

In the following, the notation $\mathbf{V}_{s,t}$ for the different views (or sub-aperture images) will be simplified as $\mathbf{V}_i$ with a bijection between (s, t) and i. The complete light field can hence be represented by a matrix $\mathbf{V} \in \mathbb{R}^{m,n}$:

$$\mathbf{V} = [vec(\mathbf{V}_1) \mid vec(\mathbf{V}_2) \mid \cdots \mid vec(\mathbf{V}_n)] \tag{1}$$

with $vec(\mathbf{V}_i)$ being the vectorized representation of the $i-$th sub-aperture image, m represents the number of pixels in each view ($m = X \times Y$) and n is the number of views in the light field.

We define here a *patch-volume* $\mathbf{p}_j \in \mathbb{R}^{q,q,n}$ to be a volumetric stack of 2D patches of dimension $q \times q$ from each of the n sub-aperture images, where j stands for the patch index. Let (x_{c_j}, y_{c_j}) be the spatial coordinates of the center pixel of the $j-$th patch. The *patch-volume* $\mathbf{p}_j$ is therefore made of a stack of 2D-patches extracted from each sub-aperture image, each 2D-patch including all pixels within the range $\left([x_{c_j} - \frac{q}{2}, x_{c_j} + \frac{q}{2}], [y_{c_j} - \frac{q}{2}, y_{c_j} + \frac{q}{2}]\right)$. Without loss of generality, the *patch-volume* will be assumed to be centered (by mean removal) in the sequel.

III. Proposed Method

Let $\mathbf{V}^L$ be the low-resolution input light field matrix and $\mathbf{V}^H$ its high-resolution version which we try to estimate. Given the observed low-resolution light field $\mathbf{V}^L$, this problem can be formulated in a Banach space as

$$\mathbf{V}^L = \Theta \mathbf{V}^H + \boldsymbol{\eta} \tag{2}$$

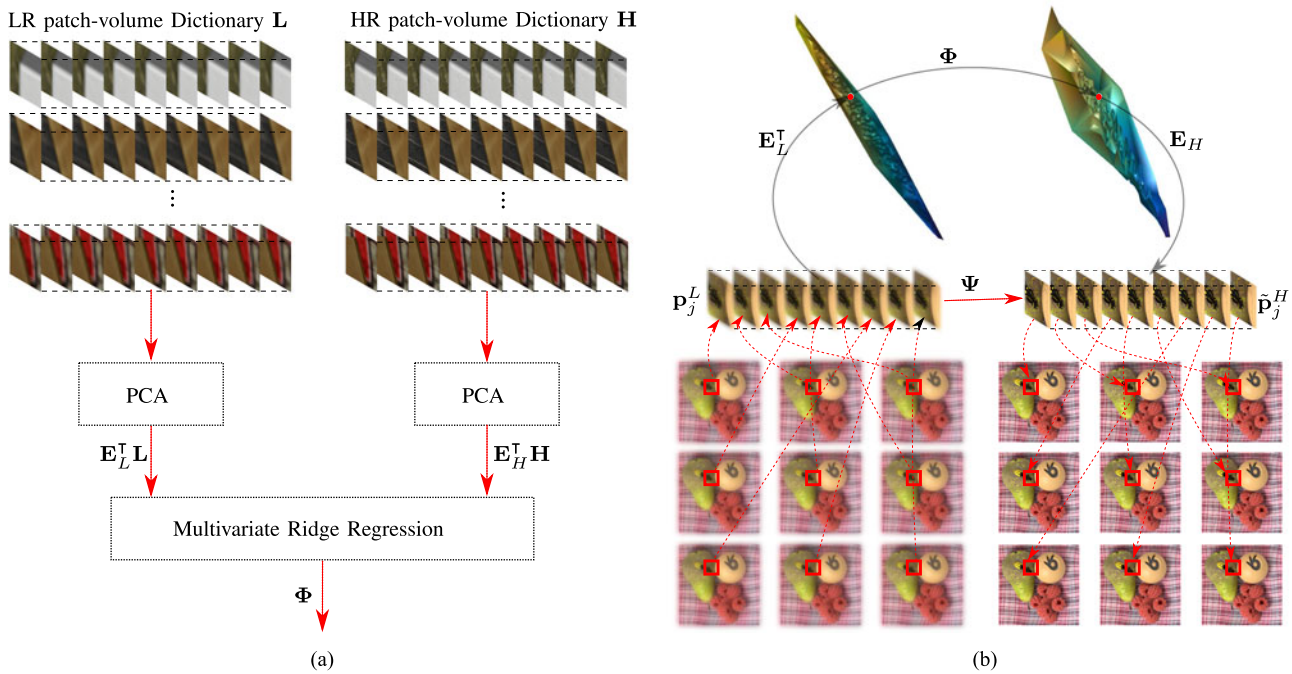


Fig. 1. Schematic diagram of the proposed spatial light field super-resolution method where (a) shows how the linear projections are learned from a set of training *patch-volumes* and (b) how the learned projections are used to restore the Light Field. The black arrows in (b) show the actual projection of the low-quality patch volume $\mathbf{p}_j^L$ onto the low-quality sub-space using $\mathbf{E}_L^\intercal$, followed by the projection from the low- to high-quality subspaces using the projection $\mathbf{\Phi}$ and then the projection from the high-quality subspace to reconstruct the high-quality patch volume $\tilde{\mathbf{p}}_j^H$. Note that the projection $\mathbf{\Psi}$ is equal to the product of these linear projection matrices as defined in (8).

where $\boldsymbol{\eta}$ is an additive noise matrix and Θ is a bounded operator composed of a blurring kernel $\mathbf{B}$ and a downsampling operator by a factor α applied on each sub-aperture image. There are many possible high-resolution light fields $\mathbf{V}^H$ which can produce the input low-resolution light field $\mathbf{V}^L$ via the acquisition model. Hence, solving this ill-posed inverse problem requires introducing some priors on $\mathbf{V}^H$, which can be a statistical prior such as a GMM model [9], or priors learned from training data as in [10] and in our proposed method.

A. *Method Overview*

Fig. 1 depicts the block diagram of the proposed spatial light field super-resolution algorithm, which consists of a Training Phase (Fig. 1 (a)) used to learn the linear projections and a Testing Phase (Fig. 1 (b)) where the linear projections are used to restore the light field. The Training phase depicted in Fig. 1 (a) is explained in more detail in sub-Section III-B and III-C. This sub-section will explain how the learned linear projections are used to restore a low-quality light field.

During Testing phase (see Fig. 1 (b)), each sub-aperture image of the unseen[1] low-quality light field $\mathbf{V}^L$ is first bi-cubic interpolated, so that both $\mathbf{V}^L$ and $\mathbf{V}^H$ have the same dimensions, and then divided into a set of overlapping patches of size $q \times q$ with an overlap of γ which, without loss of generality, was set to $\lfloor \frac{q}{3} \rfloor$. While the authors of the previous patch-based approach for super-resolving light fields [9] use a LMMSE approach with a GMM prior to estimate the 4D high-resolution patches, we instead propose an example-based method operating on 3D stacks of 2D-patches extracted from the different sub-aperture images. The stack of 2D-patches is called a *patch-volume* which is of dimension $z = q \times q \times n$. To form the stack of 2D-patches, in the sequel, we consider two cases: i)-stacking the 2D-patches which are at the same spatial location in all sub-aperture images, and ii)-stacking best matching 2D-patches across all sub-aperture images.

Example-based super-resolution methods for 2D images often rely on the assumption that the low-resolution (and respectively high-resolution 2D patches) lie on (or close to) a manifold. Neighbour embedding techniques, as in [19]–[21], were proposed for single image super-resolution. They aim at capturing the local geometry of a manifold on which the low-resolution patches are supposed to lay, by computing linear combination weights to approximate a given 2D patch from its neighbours on the manifold. The same weights are then used to reconstruct the high-resolution 2D patch. More recently, the authors in [22], [23] partition the dictionaries into multiple clusters and use a local linear mapping for each cluster to compute the weights. All these methods assume that both low- and high-resolution patches lie on two manifolds having similar geometrical structures, hence on which the patch neighborhood is preserved. In order to enforce the assumption of manifold similarity between the two distinct spaces represented by the LR and HR patches, the authors in [24] compute the weights in a subspace common to the LR and HR patches, assuming the existence of a common low-dimensional space that preserves some meaningful patch

[1]Patches from the low-quality light field are not included in the coupled dictionaries $\mathbf{L}$ and $\mathbf{H}$.

characteristics. However, the authors in [25] have shown that this assumption does not hold well because the one-to-many mapping from low- and high-resolution manifolds distort the structure of the low-resolution manifold, which negatively affects the neighbour preservation.

Here, we consider a different patch reconstruction approach, similar to the work in [16], [26], [27] which consists in learning a mapping function between the low- and high-resolution subspaces. Nevertheless, these existing methods are designed for single-image super-resolution and do not exploit the light field structure to restore the light field. Moreover, learning the mapping relation for *patch-volumes* is not a trivial task since each *patch-volume* has a considerably large dimension z which can be of the order of thousands of pixels. As the sub-aperture images correspond to different representations of the same scene captured from different viewpoints, these *patch-volumes* contain a lot of redundant information that can be exploited to reduce the dimension of the *patch-volumes* and therefore making the learning of the mapping between low- and high-resolution *patch-volumes* easier. Dimensionality reduction techniques can be used to project the low- and high- resolution *patch-volumes* onto their corresponding low-dimensional subspaces, and the learning of the mapping function can be performed between projected training *patch-volumes*.

In summary, in order to learn this mapping function, we first construct coupled dictionaries $\mathbf{L}$ and $\mathbf{H}$ of dimension $\mathbb{R}^{z,M}$ which respectively contain low- and high-resolution *patch-volumes* to be used for training. These dictionaries are not learned and are constructed by choosing M random *patch-volumes* from a set of light fields captured using the same camera model. The low-resolution *patch-volumes* are first projected onto the low-resolution subspace using the projection matrix $\mathbf{E}_L \in \mathbb{R}^{z,d_l}$ ($d_l \ll z$ is the dimension of the low-resolution subspace), and then projected onto the high-resolution subspace using a projection matrix $\mathbf{\Phi} \in \mathbb{R}^{d_h,d_l}$ ($d_h \ll z$ is the dimension of the high-quality subspace), and then projected back to the pixel domain using the projection matrix $\mathbf{E}_H \in \mathbb{R}^{z,d_h}$. In the next sections we detail the subspace projection, the mapping function learning steps and how this scheme can be extended to perform angular super-resolution.

B. Projection on Patch-Volume Subspaces

The vectorized input low-resolution *patch-volume* $\mathbf{p}_j^L$ and the vectorized high-resolution *patch-volume* $\mathbf{p}_j^H$ have a dimension z of the order of thousands of pixels. In order to reduce the dimension of the *patch-volumes*, PCA is adopted to compute the projection matrices for both low- and high-resolution *patch-volume* subspaces. The low-resolution *patch-volume* dictionary $\mathbf{L}$ and the high-resolution *patch-volume* dictionary $\mathbf{H}$ are used to compute respectively the low- and high-resolution *patch-volume* subspaces. The first step involves the computation of the respective covariance matrices

$$\mathbf{C}_L = \mathbf{L}\mathbf{L}^\intercal \qquad \mathbf{C}_H = \mathbf{H}\mathbf{H}^\intercal \tag{3}$$

The low-resolution *patch-volume* subspace projection matrix $\mathbf{E}_L$ and the corresponding high-resolution *patch-volume* subspace projection matrix $\mathbf{E}_H$ are respectively derived using the eigen-decomposition of the corresponding covariance matrices $\mathbf{C}_L$ and $\mathbf{C}_H$. Both low- and high-resolution *patch-volumes* can be projected on their corresponding subspaces whose dimension is significantly lower than the size z of the *patch-volume* while still keeping the most relevant information. One appealing property of these projections computed using PCA is that both projection matrices $\mathbf{E}_L$ and $\mathbf{E}_H$ are orthogonal and therefore one can project a *patch-volume* to and from the subspace without loosing any information *i.e.* $\mathbf{E}_L\mathbf{E}_L^\intercal = \mathbf{E}_L^\intercal\mathbf{E}_L = \mathbf{I}$ and $\mathbf{E}_H\mathbf{E}_H^\intercal = \mathbf{E}_H^\intercal\mathbf{E}_H = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Nevertheless, we use PCA to reduce the dimension of the *patch-volumes* and therefore we only use the first d_L column vectors from $\mathbf{E}_L$, and the first d_H column vectors from $\mathbf{E}_H$ based on the largest eigen-values. Reducing the number of column vectors from the projection matrices has the benefit of reducing the dimension of the *patch-volume* sub-spaces at the expense of loosing some information which may not be significant. Nevertheless, reducing the dimensionality of the problem will facilitate the learning of the mapping function between the low- and high-resolution *patch-volume* subspaces.

C. Linear Mapping Between LR and HR Subspaces

Patch-volume based super-resolution can be viewed as a regression problem *i.e.* finding a mapping function $\mathbf{\Psi} \in \mathbb{R}^{z,z}$ from the low-resolution *patch-volume* to the target high-resolution *patch-volume*. However, given that the dimension of the *patch-volume* z is generally very large, the regression learning process needs a huge amount of data ($M \gg z$) in order to learn a suitable regression model. For example, the authors in [10] employed between $1-2$ million patches to train their DCNN which takes a huge amount of time to train.

Instead, we try to learn the projection between the low- and high-resolution *patch-volume* subspaces. The major advantage of this approach is that the low- and high-resolution *patch-volume* subspaces have a dimension of d_L and d_H respectively, where both of them are significantly lower than z and therefore a smaller number of training *patch-volume* samples are needed.

We therefore use the low- and high-resolution *patch-volumes* in the coupled dictionaries, $\mathbf{L}$ and $\mathbf{H}$, which are projected respectively onto the low- and high-resolution *patch-volume* subspaces using

$$\mathbf{L}_s = \mathbf{E}_L^\intercal\mathbf{L} \qquad \mathbf{H}_s = \mathbf{E}_H^\intercal\mathbf{H} \tag{4}$$

yielding the projected coupled dictionaries $\mathbf{L}_s$ and $\mathbf{H}_s$ (see Fig. 1 (a)). The mapping function is then learned by minimizing the empirical fitting error between all the pairs of examples on the two subspaces. We consider a linear mapping function, *i.e.* we search for the projection matrix $\mathbf{\Phi}$ which minimizes the following l_2- norm regularized least squares problem

$$\mathbf{\Phi} = \arg\min_{\mathbf{\Phi}} ||\mathbf{H}_s - \mathbf{\Phi}\mathbf{L}_s||_2^2 + \lambda||\mathbf{\Phi}||_2^2 \tag{5}$$

where λ is a regularization parameter which was set to 10^{-6} in all our experiments. The regularization term is a Tikhonov regularization added to prevent singular values and provide a more stable solution than ordinary least squares. The problem is

solved using multivariate ridge regression, and the solution can be written in closed form as

$$\mathbf{\Phi} = \mathbf{H}_s \mathbf{L}_s^\intercal \left(\mathbf{L}_s \mathbf{L}_s^\intercal + \lambda \mathbf{I}\right)^{-1} \tag{6}$$

where $\mathbf{I}$ is the identity matrix. The high-resolution *patch-volume* can therefore be approximated using

$$\tilde{\mathbf{p}}_j^H = \mathbf{\Psi} \mathbf{p}_j^L \tag{7}$$

where $\mathbf{\Psi}$ is a projection matrix that can be pre-computed using

$$\mathbf{\Psi} = \mathbf{E}_H \mathbf{\Phi} \mathbf{E}_L^\intercal \tag{8}$$

The restored patch-volume $\tilde{\boldsymbol{p}}_j^H$ is de-multiplexed to restore the collocated patches within each sub-aperture image simultaneously. Overlapping regions within the *patch-volume* are averaged to minimize blocking artefacts.

The complexity to restore a *patch-volume* is of the order $O(z^2)$. In terms of execution, the proposed PCA + RR takes 158-seconds to restore the INRIA light fields while the SRCNN + LF [10] method[2] and the method of Mitra [9] are computed in 96- and 509-seconds respectively. The other methods take more than 15-minutes to restore a single INRIA light field.

D. Patch-Volume Alignment

The PCA + RR method discussed above assumes that the stack of 2D collocated patches from all sub-aperture images, which are part of the same *patch-volume*, are well aligned. However, this may not always be the case, especially when filmed using arrays of cameras capturing the scene from different viewpoints. In such systems, large disparities will exist across collocated 2D patches forming the *patch-volume*. To characterize this problem we derive a variance map $\mathbf{\Sigma}$ which is computed as

$$\mathbf{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\boldsymbol{V}_i^L - \boldsymbol{\mu}_i\right)^2 \tag{9}$$

where

$$\boldsymbol{\mu}_i = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{V}_i^L \tag{10}$$

The variance map $\mathbf{\Sigma}$ measures the variance across the n sub-aperture images. Fig. 2 shows an analysis conducted on the Buddha synthetic light field from the HCI dataset. It can be seen that the proposed PCA + RR method performs well when the pixel-variance is sufficiently low (first two rows in Fig. 2). On the other hand, the reconstructed image contains severe distortions in regions having large pixel-variance (bottom two rows in Fig. 2). Therefore, the objective here is to develop a method which reduces the variance within each *patch-volume*.

In this work, we proposed an iterative procedure where a different sub-aperture image is chosen as anchor at each iteration. The anchor sub-aperture image is used to extract overlapping 2D anchor patches. The block matching algorithm (BMA) is then used to search within a search window for the 2D patch from the other sub-aperture images which minimizes the mean square error. Therefore, the BMA method will derive those patches which

[2]The SRCNN + LF method was implemented using Caffe.

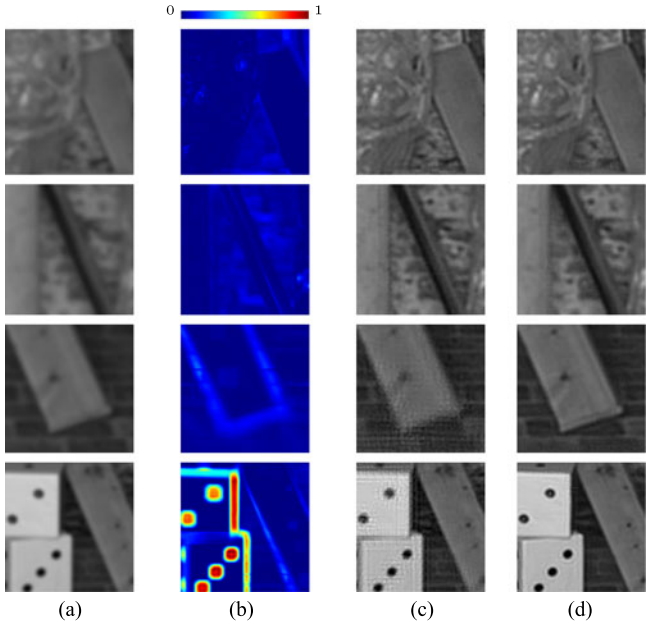


Fig. 2. Analysing crops of the center view of the Buddha light field (a) low-quality sub-aperture image, (b) variance map $\mathbf{\Sigma}$, and images super-resolved with the proposed (c) PCA + RR and (d) BM + PCA + RR methods (see Algorithm 1).

are best aligned and therefore is expected to reduce the variance within the *patch-volume*.

At each iteration, all pixels within the anchor sub-aperture image are restored. However, the remaining sub-aperture images will contain pixels which have not been super-resolved (not considered when forming the *patch-volumes* by the block matching search). The sub-aperture image with the largest number of unprocessed pixels (holes) is selected as the new anchor sub-aperture image for the next iteration. This process terminates once all pixels within the light field have been processed. The BM + PCA + RR algorithm is summarized in Algorithm 1 and a demo will be made available upon publication.

In order to characterize the variance within each *patch- volume*, we define the *patch-volume variance* $\boldsymbol{\sigma}_j$ which is computed using

$$\boldsymbol{\sigma}_j = \frac{1}{q^2} \sum_{j} \mathbf{\Sigma}_j \tag{11}$$

which is the average variance within the $j-$th $q \times q$ patch from $\mathbf{\Sigma}$. Fig. 3 shows the distribution of $\boldsymbol{\sigma}_j$ with and without block matching for the Buddha and Still Life light fields. It can be immediately noticed that block matching manages to significantly reduce the *patch-volume variance*. It can also be seen from the last two columns of Fig. 2 that reducing the *patch-volume variance* significantly improves the quality of the super-resolved light field, especially in regions with high disparity.

E. Angular Super-Resolution

A simple way to produce a novel view between two neighbouring sub-aperture images is to apply bilinear interpolation. However, such approach loses high frequency information and

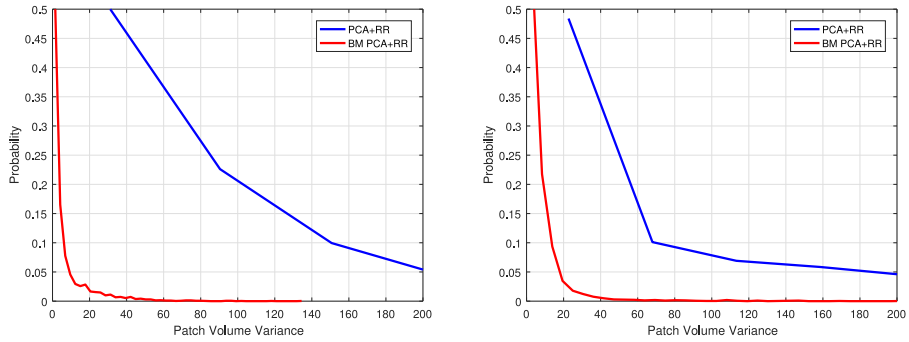


Fig. 3. Probability distribution of the *patch-volume variance* $\boldsymbol{\sigma}_j$ for the (a) Buddha and (b) Stilllife light fields.

Algorithm 1: BM + PCA + RR Light field super-resolution algorithm.

$\tilde{\boldsymbol{V}}^H$ = function bm_pca_rr_LFSR $(\boldsymbol{V}^L, \boldsymbol{\Psi})$
Initialize: Set the center view as the anchor sub-aperture image *i.e.* $i_A = \lceil \frac{n}{2} \rceil$, where n is the number of sub-aperture images
while *not all pixels of the light field are reconstructed* **do**
  for $j = 1 : P$ **do**
    Choose the $j-$th 2D patch, denoted by $\boldsymbol{p}^L_{i_A,j}$, from the anchor sub-aperture image and put it as the i_A-th patch in the *patch-volume* $\boldsymbol{p}^L_j$.
    for $i = 1 : N$ **do**
      if $i \neq i_A$ **then**
        Use the block-matching algorithm to find the patch within the $i-$th sub-aperture image which is closest to $\boldsymbol{p}^L_{i_A,j}$ in terms of mean square error and put it as the $i-$th patch in the *patch-volume* $\boldsymbol{p}^L_j$.
      end
    end
    Project the low-quality patch-volume $\boldsymbol{p}^L_j$ using (7) to synthesize the high-quality *patch-volume* $\tilde{\boldsymbol{p}}^H_j$.
    Demultiplex the reconstructed patch-volumes to reconstruct the high quality light field, where overlapping pixels are averaged.
  end
  Compute the number of holes (pixels which have not been selected by BMA) in each sub-aperture image and set the image with the largest number of holes as the new anchor sub-aperture image.
end

produces sub-aperture images that are blurred. On the other hand, the authors in [8] assume that accurate depth maps are used as a geometric guidance to perform angular SR. However, as shown in [10], this method fails in practice because obtaining reliable depth-maps from a light field is prone to errors which significantly reduces its performance.

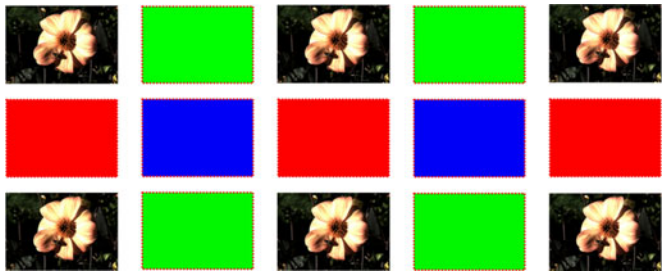


Fig. 4. Schematic diagram of the proposed angular light field super-resolution method. The green boxes represent case 1), red boxes represent case 2) and blue boxes represent case 3). Note that the green, red and blue boxes are empty during testing phase, but these sub-aperture images are available in $\mathbf{H}$ during training.

The DCNN based scheme in [10] exploits the local neighbourhood, but tries to approximate a global model for all angular viewpoints. However, deriving the angular super-resolution of a light field using one global function generally provides poor performance. This can be explained by the fact that each sub-aperture image represents the same scene from different viewpoints, and therefore a global interpolation function suitable for all different angular displacements will provide sub-optimal performance. Instead, we employ a new method where a linear interpolation function for each missing sub-aperture image is learned as shown in Fig. 4. Similar to the work of Yoon *et. al.* [10], our approach is a data driven supervised learning method. However, in our case we use multivariate ridge regression to learn the interpolation function for each unknown view-point (s,t). Moreover, while one interpolation function is learned for every missing sub-aperture image, the proposed angular super-resolution algorithm is still a learning patch-based strategy. We here define three cases to learn a mapping function:

1) The case where sub-aperture images at locations $(s-1,t)$ and $(s+1,t)$ are available. In this case we extract the 2D patches at locations $(s-1,t)$, (s,t) and $(s+1,t)$ from the *patch-volumes* contained in the high-quality dictionary $\mathbf{H}$. All the 2D patches at location $(s-1,t)$ and $(s+1,t)$ are concatenated to form the dictionary of the known part $\mathbf{H}^k_i$ while the dictionary of unknown part $\mathbf{H}^u_i$ is formed using all 2D patches at location (s,t).

2) The case where sub-aperture images at locations $(s, t-1)$ and $(s, t+1)$ are available is very similar to the above case. This time the 2D patches from $\mathbf{H}$ at location $(s, t-1)$ and $(s, t+1)$ are concatenated to form the dictionary of known part $\mathbf{H}_i^k$ while the 2D patches at location (s, t) are used to construct the dictionary of unknown part $\mathbf{H}_i^u$.
3) The case where the neighbouring horizontal and vertical sub-aperture images are missing. In this case, the 2D patches from $\mathbf{H}$ at locations $(s-1, t-1)$, $(s-1, t+1), (s+1, t-1)$ and $(s+1, t+1)$ are concatenated to construct the dictionary of the known part $\mathbf{H}_i^k$ while the 2D patches at location (s, t) are used to construct the dictionary of unknown part $\mathbf{H}_i^u$.

The projection matrix $\boldsymbol{\beta}_i$ is computed for every unknown sub-aperture image i, and is used to estimate those sub-aperture images which need to be interpolated. Therefore, for a 5×5 matrix of sub-aperture images, we need 56 $\boldsymbol{\beta}_i$ projection matrices (one for every missing sub-aperture image) to approximate the missing sub-aperture images to increase the angular resolution to 9×9. Therefore, unlike the DCNN scheme, our method models the angular super-resolution problem using multi-linear models, one for every missing view. The projection matrix is approximated using multivariate ridge regression using its closed form solution

$$\boldsymbol{\beta}_i = \mathbf{H}_i^u \mathbf{H}_i^{k\intercal} \left(\mathbf{H}_i^u \mathbf{H}_i^{u\intercal} + \lambda \mathbf{I}\right)^{-1} \tag{12}$$

where λ is a regularization parameter set to 10^{-6}, and $\boldsymbol{\beta}_i$ is the interpolation matrix for the $i-$th unknown sub-aperture image. The projection matrix for each unknown sub-aperture image $\boldsymbol{\beta}_i$ can be pre-computed.

During the testing phase, the input light field is first up-scaled in the angular dimension by inserting empty sub-aperture images that need to be restored using the proposed method. Therefore, each *patch-volume* $\mathbf{p}_j$ contains a number of empty 2D patches which need to be approximated. The neighbourhood of the empty 2D patches $\mathbf{p}_j(s, t)$ within the *patch-volume* are used to determine which of the above mentioned cases will be used to construct the vector $\boldsymbol{\Gamma}_j$. In case the horizontal neighbours are available, $\boldsymbol{\Gamma}_j$ is constructed by concatenating the vectorized representations of the 2D patches $\mathbf{p}_j(s-1, t)$ and $\mathbf{p}_j(s+1, t)$. Otherwise, if the vertical neighbours are available, $\boldsymbol{\Gamma}_j$ is constructed by concatenating the vectorized representations of the 2D patches $\mathbf{p}_j(s, t-1)$ and $\mathbf{p}_j(s, t+1)$. The last case is when neither of the vertical and horizontal neighbours are available, in which case the vectorized representation of $\mathbf{p}_j(s-1, t-1)$, $\mathbf{p}_j(s-1, t+1), \mathbf{p}_j(s+1, t-1)$ and $\mathbf{p}_j(s+1, t+1)$ are concatenated to construct $\boldsymbol{\Gamma}_j$. The unknown patch is then approximated using

$$\tilde{\mathbf{p}}_j(s, t) = \boldsymbol{\beta}_i \boldsymbol{\Gamma}_j \tag{13}$$

In this work we only consider collocated patches since only local neighbouring 2D patches are used, where the disparities are expected to be relatively low.

IV. Experimental Results

The experiments conducted in this paper use both synthetic and real-world light fields from publicly available datasets. Here we use the synthetic light fields of HCI[3] for parameter selection and also to get a better understanding of the performance of the proposed method. It was also tested on two real-world light fields captured using different technologies: i) the Stanford dataset[4] captured using a moving camera mounted on a gantry, where the captured light fields generally have larger disparities across sub-aperture images and ii) the INRIA dataset[5] which was captured using micro-lens technology (a Lytro Illum camera) with lower disparities across sup-aperture images. While the sub-aperture images of the Stanford data set are available, the light fields from the INRIA database were decoded using the method in [28] as mentioned in the website.

In all the experiments we considered a 9×9 array of sub-aperture images from each dataset. For computational purposes, the high resolution images of the synthetic HCI light fields were downscaled to 384×384 per sub-aperture image. The Stanford light fields have different resolutions, but the high resolution sub-aperture images were scaled such that the lowest dimension was set to 400 pixels while maintaining the original aspect ratio. On the other hand, the original light field resolutions were used for the INRIA dataset *i.e*. 625×434. In all the experiments involving spatial super-resolution, each sub-aperture image was blurred using a Gaussian filter of size 7×7 with standard deviation of 1.6 followed by downscaling by a factor of α, where unless otherwise specified is set to $\alpha = 2$.

In every experiment we use a leave-one out methodology to train the projection matrix $\boldsymbol{\Psi}$ for every dataset considered *i.e*. if we take the Buddha light field from the HCI dataset as an example, we extract 2000 *patch-volumes* from all the other light fields of the same dataset except for the Buddha. This was done to learn the mapping function $\boldsymbol{\Psi}$ for the specific camera model, without of course biasing the results by including *patch-volumes* from the light field under test. The total number of patch-volumes varied depending on the number of light fields considered for each dataset: i) 16 K for the HCI dataset ii) 22 K for the Stanford dataset iii) 14 K for the INRIA dataset. While we observed that better performance can be achieved using a larger set of training *patch-volumes*, it did not significantly improve when using dictionaries larger than the above mentioned sizes.

The performance of the proposed method was compared to the single-image super-resolution schemes [14]–[16], where these methods are applied on every sub-aperture image. It was also compared against two light-field super-resolution methods including the work of Mitra *et. al*. [9] and the DCNN based scheme [10]. The angular super-resolution method was also compared against the recent view synthesis method published in [12]. In all the experiments we used the code and pre-trained models provided by the authors or are publicly available on-line.

[3]HCI dataset: https://hci.iwr.uni-heidelberg.de/hci/softwares/light_field_analysis
[4]Stanford dataset: http://lightfield.stanford.edu/
[5]INRIA dataset: http://www.irisa.fr/temics/demos/lightField/CLIM/DataSoftware.html

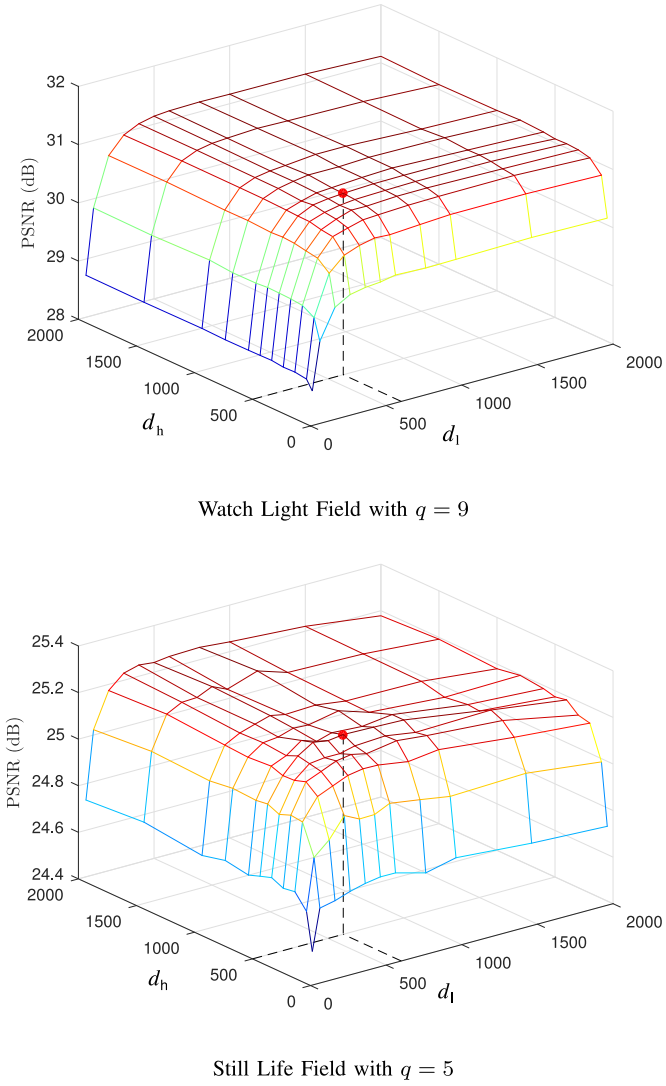


Fig. 5. PSNR analysis computed by using different light fields from the HCI dataset and different patch sizes.

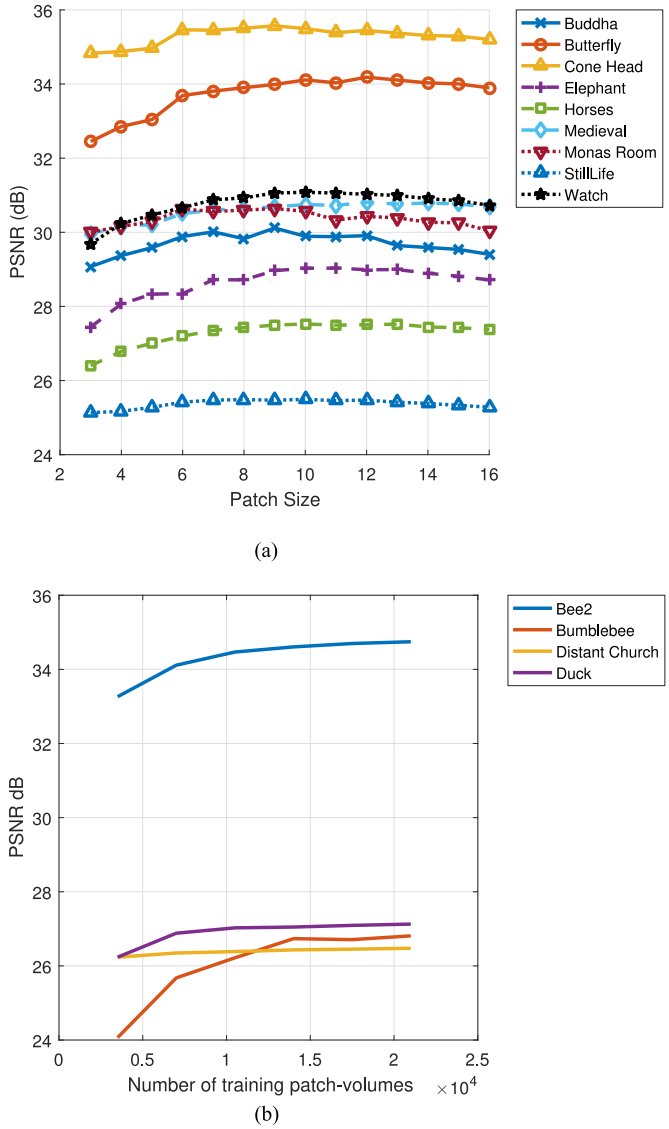


Fig. 6. PSNR analysis as a function of (a) patch size and (b) number of training *patch-volumes* for different light fields.

A. *Parameter Selection*

The method presented in this paper has four parameters that have to be tuned: i) the dimension of the low-resolution *patch-volume* subspace d_l, ii) the dimension of the high-resolution *patch-volume* subspace d_h, and iii) the patch size q and iv) the number of *patch-volumes*. Fig. 5 shows the performance in terms of PSNR using the PCA + RR method obtained by fixing the patch size q and varying the dimension of the subspaces. Similar plots were obtained using different light fields and different patch sizes but due to space constraints could not be included in this paper. It can be seen that for different patch sizes and light fields the graphs seem to saturate and that good performance is achieved when setting $d_l = d_h = 500$.

The optimal patch size q is derived using the PSNR analysis illustrated in Fig. 6(a) where we set $d_l = d_h = 500$ and use nine light fields from the HCI dataset. It can be seen that the performance starts increasing with increasing patch sizes and it saturates (or starts degrading) at $q > 10$. Therefore setting $q = 10$ seems to be an optimal parameter for the synthetic light fields. Unless otherwise specified, we will use these parameters for the following simulations. In order to assess the effect of the number of training *patch-volumes* we super-resolve four INRIA light fields with a magnification factor of 2. The results in Fig. 6 (b) show that higher PSNR is achieved when using more training *patch-volumes*. Nevertheless, the performance does not significantly improve when using more than 14K *patch-volumes* for training.

B. *Spatial Super-Resolution of Synthetic Light Fields*

In this section we analyse the performance of the proposed method using the synthetic light fields of the HCI dataset. In these experiments we set the magnification factor $\alpha = 2$ and compare our approach to state-of-the-art methods in the area of single-image super-resolution [14]–[16] and light field

TABLE I
PSNR ANALYSIS OF SPATIAL SUPER-RESOLUTION ON THE HCI DATASET

Light Field Name	Bicubic	Dong [15]	Peleg [14]	Zhang [16]	Mitra [9]	SRCNN + LF [10]	PCA + RR	BM PCA + RR
Buddha	27.2387	27.1295	27.5482	27.6509	22.2204	27.3565	29.9276	**31.1277**
Butterfly	29.7868	29.5468	30.0613	30.1555	28.8319	29.2139	**33.7600**	33.5527
Cone Head	34.2545	34.1121	34.3921	34.5577	29.5444	31.3876	35.3642	**36.6109**
Horses	24.5134	24.6663	24.8337	24.8801	23.9871	24.6014	**27.9206**	27.5649
Medieval	29.1129	29.0211	29.2461	29.3695	28.3404	28.6737	**31.1431**	31.0744
Monas Room	27.3276	27.1595	27.6295	27.7423	23.5765	27.1786	30.2960	**30.7295**
Elephant	24.4555	24.5238	24.7661	24.7881	22.8509	24.1461	**29.5462**	28.3563
Watch	25.9329	25.6987	26.2238	26.2930	23.9491	25.5008	**30.9749**	29.9094
StillLife	24.2778	24.0777	24.4509	24.5284	21.4812	23.8061	26.0804	**26.1894**

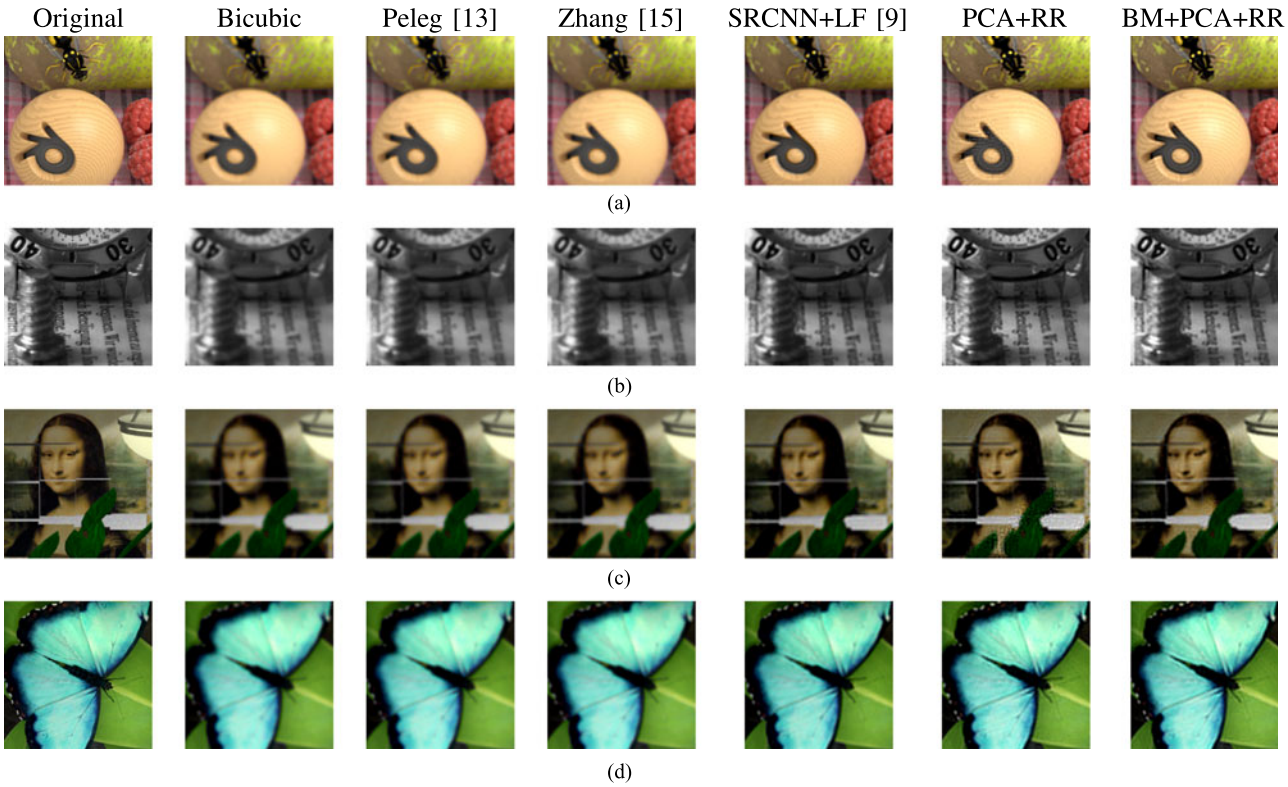


Fig. 7. Cropped regions of the central view obtained using different spatial super-resolution techniques obtained using the synthetic HCI dataset. (a) Still life light field. (b) Watch light field. (c) Monas room light field. (d) Butterfly light field.

super-resolution [9], [10]. The results in Table I and Fig. 7 clearly show the superiority of the proposed method where it achieves PSNR gains of up to 4 dB over the second best algorithm. It can be seen, especially from the images in Fig. 7, that single-image super-resolution schemes and bi-cubic interpolation generally provide blurred sub-aperture images. On the other hand, both the recent deep learning based scheme and our proposed method generate images which are sharper, with our proposed BM + PCA + RR providing the sharpest and most accurate results. Unlike existing schemes, our proposed method exploits the redundant information present within the light-field structure through the use of *patch-volumes* and low-dimensional sub-space projections. Moreover, this gain over DCNN is achieved using a significantly smaller amount of training data which can be efficiently trained and pre-computed using a normal computer without the need of GPUs to speed up the learning process. These results also show that PCA + RR performs best with light fields like Watch and Butterfly which have low disparity while BM + PCA + RR performs better with light fields having larger disparities such as Buddha and Stillife. Moreover, unlike existing schemes, our method ensures that the reconstructed sub-aperture images are coherent across all sub-aperture images by using of *patch-volumes* and therefore exploiting the light field structure.

C. Spatial Super-Resolution of Real World Light Fields

In this experiment we use the light field images provided by Stanford which are captured using a camera mounted on a gantry. The results in Table II clearly show that most of the

TABLE II
PSNR ANALYSIS OF SPATIAL SUPER-RESOLUTION (MAGNIFICATION FACTOR $\times 2$) USING THE STANFORD LIGHT FIELD IMAGES

Light Field Name	Bicubic	Dong [15]	Peleg [14]	Zhang [16]	Mitra [9]	SRCNN + LF [10]	PCA + RR	BM PCA + RR
Amethyst	30.0753	29.8729	30.3426	30.4042	27.5084	30.1516	33.2770	**33.3482**
Bracelet	26.2049	26.1827	26.5485	26.5931	19.0219	25.4509	25.1423	**28.4921**
Bunny	31.2884	30.8454	31.5553	31.6312	27.4327	29.4336	34.8505	**36.8623**
Chess	29.8041	29.4868	30.1000	30.1758	24.8064	29.0012	32.9157	**33.4924**
Eucalyptus Flowers	30.4913	30.3423	30.7207	30.7922	29.1275	30.5894	33.0552	**33.1837**
Jelly Beens	40.2994	39.2555	40.4947	**40.6129**	26.9460	29.7700	40.5246	38.8378
Lego Bulldozer	26.0729	25.8941	26.4372	26.4999	20.2748	26.4719	26.8637	**29.0056**
Lego Gantry Self Portrait	27.1717	27.0497	27.5291	27.6457	20.3284	27.4146	28.5695	**28.8770**
Lego Knights	26.8735	26.7013	27.2120	27.2896	19.7672	26.0193	28.0270	**29.6041**
Lego Truck	30.4957	30.3576	30.7638	30.8524	28.6439	30.8273	**33.6306**	33.5345
Treasure	25.8466	25.7149	26.0605	26.1300	21.8995	26.4346	27.6640	**28.2319**

TABLE III
PSNR ANALYSIS OF SPATIAL SUPER-RESOLUTION (MAGNIFICATION FACTOR $\times 2$) USING THE INRIA LIGHT FIELD IMAGES

Light Field Name	Bicubic	Dong [15]	Peleg [14]	Zhang [16]	Mitra [9]	SRCNN + LF [10]	PCA + RR	BM + PCA + RR
Bee2	30.5666	30.3450	30.9149	30.3152	25.1065	29.8795	**34.6064**	33.5067
Bumblebee	26.6509	26.4444	27.0462	26.4769	17.5138	27.6928	26.7346	**28.7119**
DistantChurch	24.3331	24.3298	24.5979	24.4754	24.3143	25.1071	26.4316	**26.4920**
Duck	23.5316	23.5510	23.8874	23.7356	21.4510	24.8502	**27.0484**	26.3975
Framed	27.5230	27.3720	27.9161	27.7436	25.4695	27.9237	**30.8640**	30.4825
Fruits	28.4813	28.4310	28.8230	28.6050	22.7371	29.5817	31.6108	**32.0372**
Mini	27.6342	27.6546	27.9222	27.7356	25.8211	28.2645	**30.8350**	30.2396
Rose	33.6350	33.6134	33.8568	33.8966	28.8269	34.4949	**37.4667**	36.9080

TABLE IV
PSNR ANALYSIS OF SPATIAL SUPER-RESOLUTION USING THE INRIA LIGHT FIELD IMAGES AT DIFFERENT MAGNIFICATION FACTORS

Light Field Name	$\times 3$					$\times 4$				
	Bicubic	ANR [29]	NCSR [30]	PCA + RR	BM + PCA + RR	Bicubic	ANR [29]	NCSR [30]	PCA + RR	BM + PCA + RR
Bee2	29.6867	30.6162	30.5056	**31.7366**	31.2349	28.4433	**29.2275**	28.3437	29.1141	28.9283
Bumblebee	25.5401	26.4189	24.7068	22.8932	**26.4321**	24.1508	**24.7748**	21.7147	20.4922	24.0889
DistantChurch	23.7253	24.1354	24.2603	24.5221	**24.5621**	22.9630	23.1522	22.5218	**23.1835**	23.1028
Duck	23.0798	23.7197	23.6979	**24.1611**	23.9780	22.1628	**22.5774**	20.4905	22.0103	21.9914
Framed	26.7072	27.4263	27.3974	28.1060	**28.1600**	25.5819	26.0354	25.8239	26.1740	**26.1741**
Fruits	27.7360	28.4712	27.9654	28.6624	**29.3045**	26.6765	**27.1361**	25.5000	26.3563	26.9609
Mini	26.9273	27.4880	27.0742	**28.8523**	28.4511	26.0561	26.3788	26.4890	**27.1800**	26.8443
Rose	32.6970	33.3592	33.0670	**34.8321**	34.5860	31.538	31.8898	31.0761	**32.6694**	32.5134

time it is the BM + PCA + RR method which outperforms all other schemes including PCA + RR. This is mainly attributed to the fact that some of the Stanford light fields have large disparities, and the alignment using block-matching improves the performance of the reconstructed light fields by reducing the inter-patch variance.

Table III shows the average PSNR of the INRIA light fields using different super-resolution techniques. It can be seen that the proposed methods again outperform all the state-of-the-art methods considered in this work with PCA + RR having the best overall performance. This might be attributed to the fact that the light fields captured with the micro-lens technology have smaller disparities and therefore alignment of the *patch-volumes* does not improve performance. It can be mentioned at this stage that the performance of Mitra *et al.* [9], is significantly lower than the bi-cubic interpolated light fields. This confirms the results presented in [10] where similar poor performance were shown for both methods in [8], [9] which exploit depth information to perform light field super-resolution. The results in Table IV further demonstrate the performance of both proposed schemes at larger magnification factors ($\times 3$, $\times 4$).

The subjective evaluation in Fig. 8 shows the results on real-world light fields. In this experiment we converted each image into the YC_bC_r colorspace, where the luminance component is restored using the following algorithms while the chrominance channels were up-scaled using bicubic interpolation. Similar to the above results, the method of Peleg *et al.*. [14] provides blurred sub-aperture images. The DCNN based method in [10] gives improved performance over single image super-resolution techniques. Nevertheless, our proposed BM +

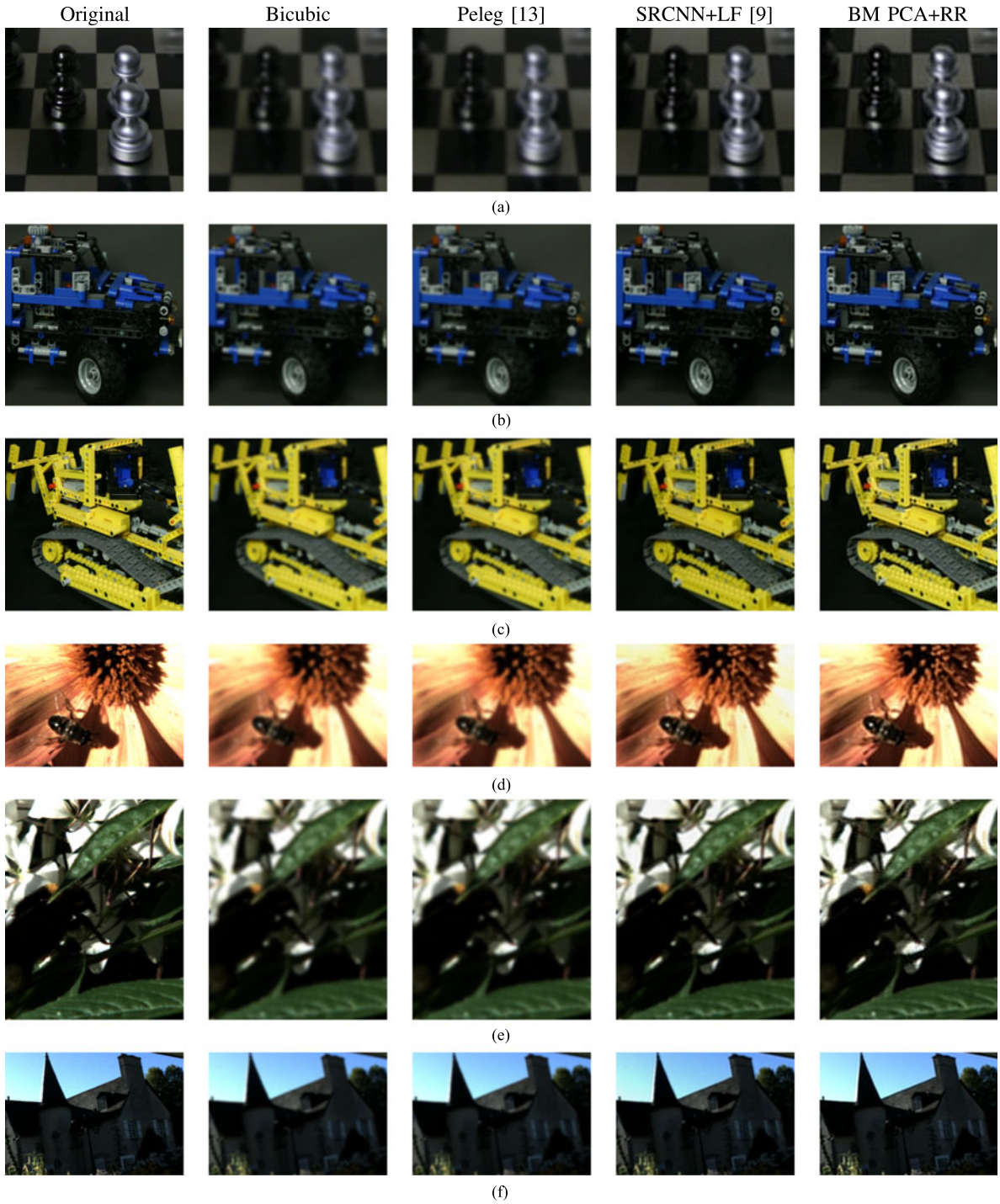


Fig. 8. Central views spatially super-resolved using different methods on the real light fields obtained from the Stanford (a–c) and INRIA (d–f) datasets. (a) Chess Light field. (b) Lego Truck Light field. (c) Lego Bulldozer Light field. (d) Bee2 Light field. (e) Bumblebee Light field. (f) Framed Light field.

PCA + RR method yields sub-aperture images with better texture detail.

D. Angular Super-Resolution

In these experiments we reduce the angular resolution by removing even column and row sub-aperture images, resulting in a matrix of 5×5 high-resolution and undistorted sub-aperture images. Angular super-resolution methods try to approximate novel sub-aperture images to increase the angular resolution of the light field. We compare our proposed angular super-resolution method with bilinear interpolation between neighbouring sub-aperture images, the SRCNN + LF method in [10] and the view synthesis method presented in [12]. The angularly super-resolved views $(6, 6)$ are shown in Fig. 9, which corresponds to case 3 described in Section III-E. It can be

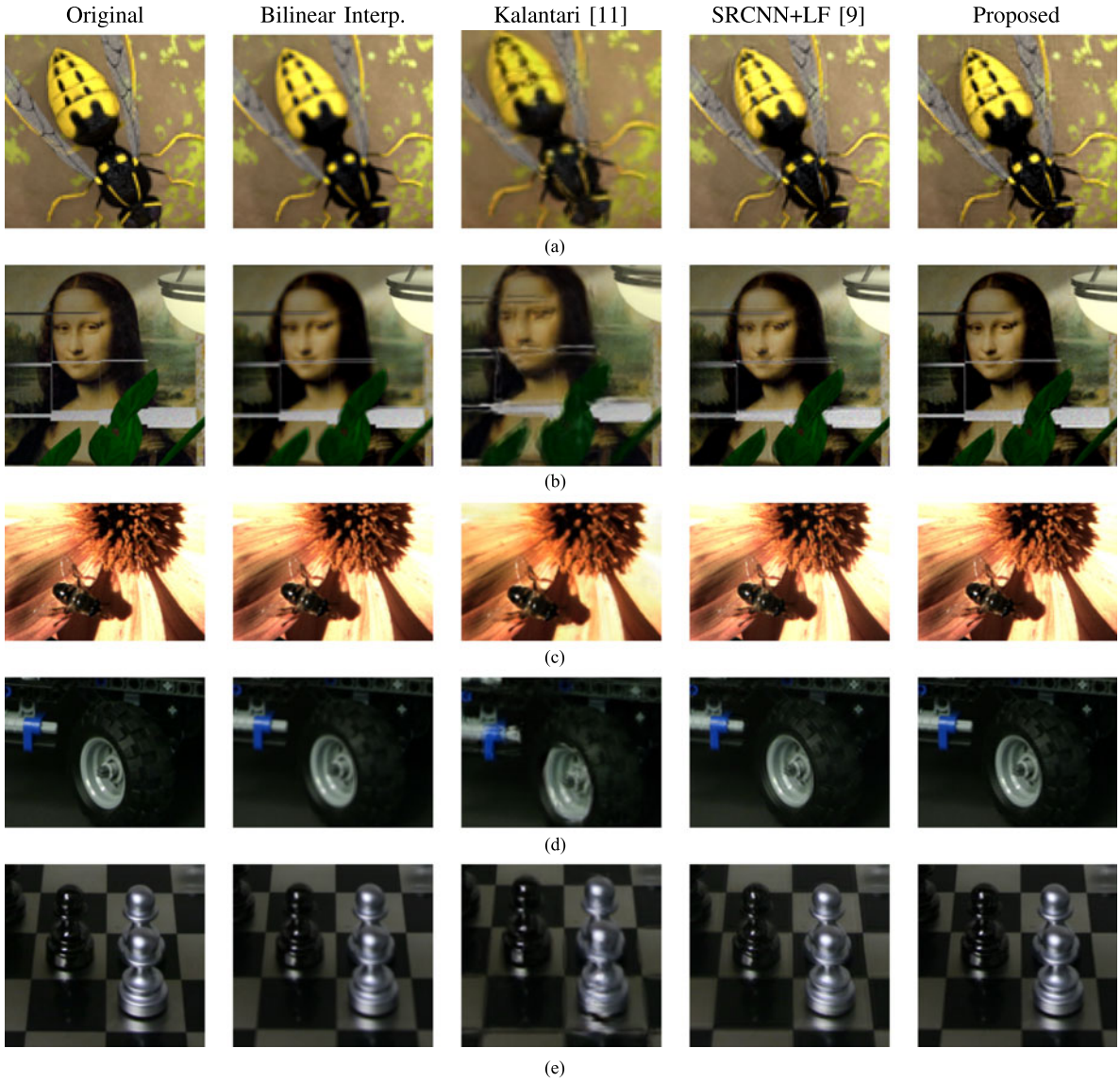


Fig. 9. Angular super-resolution analysis at interpolated sub-aperture image at (6, 6) which has neither horizontal nor vertical sub-aperture neighbours. These images are obtained using the HCI (a, b), INRIA (c), and Stanford (d, e) datasets. (a) Still life Light field. (b) Mona.s Room Light field. (c) Bee2 Light field. (d) Lego Truck Light field. (e) Chess Light field.

noticed that the images produced using simple bilinear interpolation are generally blurred, especially when considering the synthetic light fields and the light fields provided by Stanford which have larger disparities and thus harder to approximate using this simple method. The view synthesis method presented in [12] manages to restore the light fields captured by the Illum camera relatively well, however it fails to restore the other light fields (HCI and Stanford), mainly because the model has been learned on Illum light fields and should be retrained to test the method on other types of light fields. The other deep learning method of Yoon *et al*. [10] manages to recover some texture detail. However, the reconstructed interpolated sub-aperture images are generally noisy. Our proposed angular super-resolution approach provides sharper images closer to the ground truth and these improvements are more visible in synthetic and on images obtained from the Stanford dataset.

V. Spatial-Angular Super-Resolution

In the final experiment we analyse the combined spatial and angular super-resolution and compare it with the deep learning method presented in [10]. The 9×9 matrix of sub-aperture images is first angularly down-sampled by removing even rows and columns and then down-scaled by a factor of 2. Fig. 10 compares subjectively the results obtained using the state-of-the-art deep learning based method proposed in [10] with our combined scheme (spatial super-resolution using BM + PCA + RR and using the proposed angular super-resolution (Section III-E) method to restore the original angular resolution). One can notice that our proposed method manages to restore higher quality sub-aperture images both spatially (at angular resolution (5, 5) which is available in the 5×5 matrix) and angularly (6, 6) (which is angularly super-resolved). These results, and all results presented in the previous sections

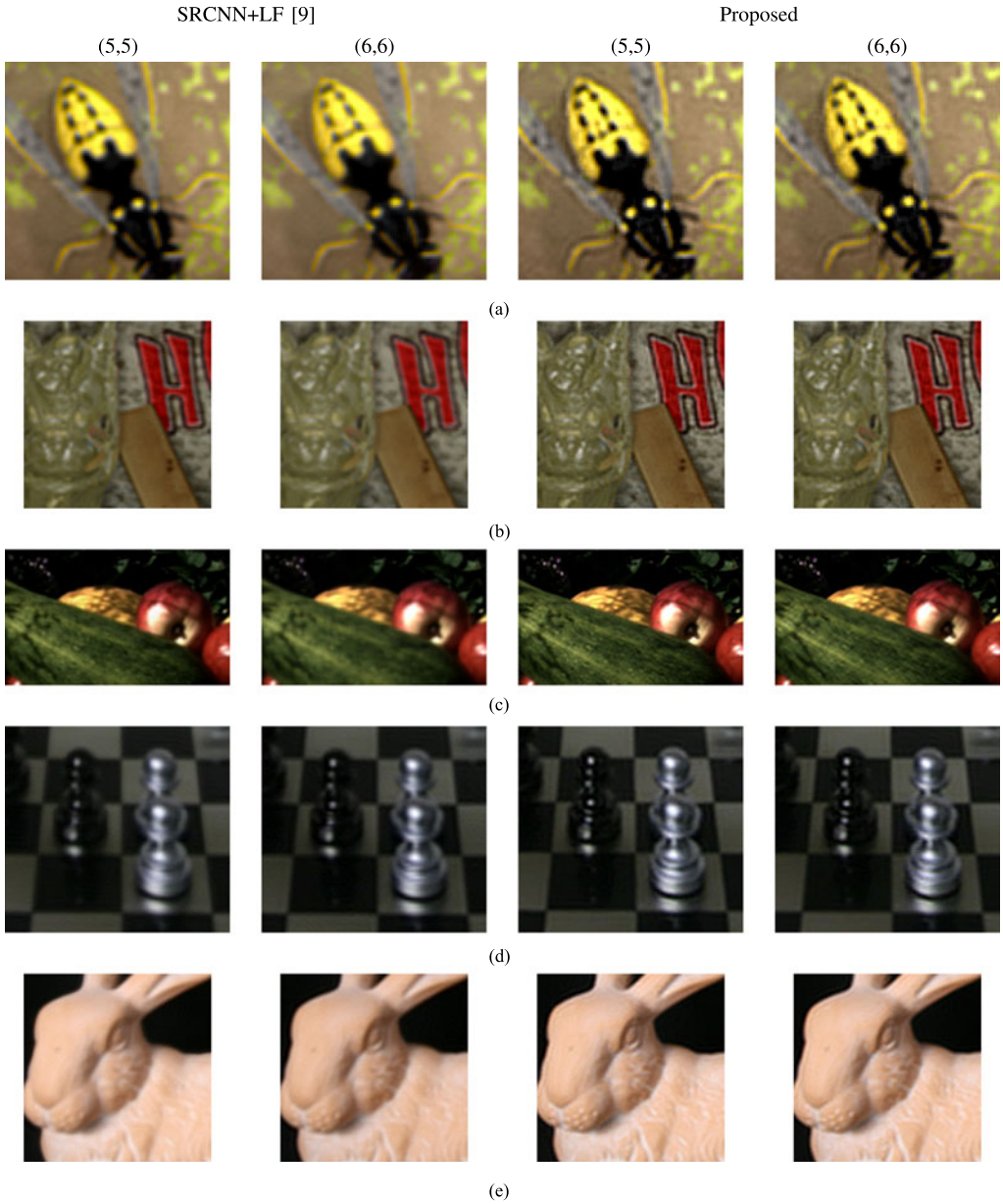


Fig. 10. Spatial and angular super-resolution analysis at spatially super-resolved center view (5, 5) and interpolated sub-aperture image at (6, 6) which has neither horizontal nor vertical sub-aperture neighbours. These images are obtained using the HCI (a, b), INRIA (c), and Stanford (d, e) datasets.

demonstrate that by properly exploiting the light field structure, linear models can effectively outperform more complicated deep learning based approaches for both angular and spatial super-resolution. Supplementary multimedia files uploaded on ScholarOne show that the super-resolved light fields (spatial and angular SR) using our proposed method are more coherent across sub-aperture images for real-world light fields (less flickering across sub-aperture images) and provides superior refocused images when compared to the DCNN scheme presented in [10].

VI. Conclusion

In this paper, we have proposed a novel super-resolution algorithm for light fields. The method is shown to outperform recent techniques based on deep convolutional neural networks for both spatial and angular super-resolution. In addition, the approach is simple, and does not require millions of training samples to learn the projections between high- and low-resolution subspaces and to increase the angular resolution in contrast to prior solutions based on deep convolutional networks. Experimental results show significant improvement over the state-of-the-art spatial super-resolution schemes, achieving average PSNR gains of 3.7dB, 4.2dB and 2.1dB on the HCI, Stanford and INRIA datasets respectively. Moreover, subjective results clearly demonstrate that the proposed spatial and angular super-resolution method outperforms more complex schemes based on deep learning. Future work will involve to learn dictionaries of *patch-volumes* that are more representative and to learn multiple linear mappings.

REFERENCES

[1] R. Ng, "Digital light field photography," Ph.D. dissertation, Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, 2006.

[2] B. Wilburn *et al.*, "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, Jul. 2005.

[3] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.

[4] F. Nava and J. Luke, "Simultaneous estimation of super-resolved depth and all-in-focus images from a plenoptic camera," in *Proc. 3DTV Conf.*, 2009, pp. 1–4.

[5] D. Cho, M. Lee, S. Kim, and Y.-W. Tai, "Modeling the calibration pipeline of the Lytro camera for high quality light-field image reconstruction," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 3280–3287.

[6] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution as sparse representation of raw image patches," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 3280–3287.

[7] T. E. Bishop, S. Zanetti, and P. Favaro, "Light field superresolution," in *Proc. Intell. Comput. Commun. Process.*, 2009, pp. 1–9.

[8] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 36, no. 3, pp. 606–619, Mar. 2014.

[9] K. Mitra and A. Veeraraghavan, "Light field denoising, light field superresolution and stereo camera based refocusing using a GMM light field patch prior," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop Comput. Cameras Displays*, 2012, pp. 22–28.

[10] Y. Yoon, H. G. Jeon, D. Yoo, J. Y. Lee, and I. S. Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Dec. 2015, pp. 57–65.

[11] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199. [Online]. Available: http://dx.doi.org/10.1007/978–3-319-10593-2_13

[12] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 193:1–193:10, Nov. 2016. [Online]. Available: http://doi.acm.org/10.1145/2980179.2980251

[13] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 46:1–46:12, Jul. 2013. [Online]. Available: http://doi.acm.org/10.1145/2461912.2461914

[14] T. Peleg and M. Elad, "A statistical prediction model based on sparse representations for single image super-resolution," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2569–2582, Jun. 2014.

[15] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[16] K. Zhang, B. Wang, W. Zuo, H. Zhang, and L. Zhang, "Joint learning of multiple regressors for single image super-resolution," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 102–106, Jan. 2016.

[17] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. Annu. Conf. Comput. Grap. Interact. Tech.*, 1996, pp. 31–42. [Online]. Available: http://doi.acm.org/10.1145/237170.237199

[18] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. 23rd Annu. Conf. Comput. Grap. Interact. Tech.*, 1996, pp. 43–54. [Online]. Available: http://doi.acm.org/10.1145/237170.237200

[19] H. Chang, D. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2004, pp. 275–282.

[20] W. Fan and D. Yeung, "Image hallucination using neighbor embedding over visual primitive manifolds," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–7.

[21] T. Chan, J. Zhang, J. Pu, and H. Huang, "Neighbor embedding based super-resolution algorithm through edge detection and feature selection," *Pattern Recognit. Lett.*, vol. 30, no. 5, pp. 494–502, Apr. 2009.

[22] K. Zhang, D. Tao, X. Gao, X. Li, and Z. Xiong, "Learning multiple linear mappings for efficient single image super-resolution," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 846–861, Mar. 2015.

[23] J. C. Ferreira, E. Vural, and C. Guillemot, "Geometry-aware neighborhood search for learning local models for image superresolution," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1354–1367, Mar. 2016.

[24] X. Gao, K. Zhang, D. Tao, and X. Li, "Joint learning for single-image super-resolution via a coupled constraint," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 469–480, Feb. 2012.

[25] K. Su, Q. Tian, Q. Xue, N. Sebe, and J. Ma, "Neighborhood issue in single-frame image super-resolution," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2005.

[26] J. S. Choi and M. Kim, "Single image super-resolution using global regression based on multiple local linear mappings," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1300–1314, Mar. 2017.

[27] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. A. Morel, "Single-image super-resolution via linear mapping of interpolated self-examples," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5334–5347, Dec. 2014.

[28] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1027–1034.

[29] R. Timofte, V. De, and L. V. Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1920–1927.

[30] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1620–1630, Apr. 2013.



Reuben A. Farrugia (S'04–M'09) received the first degree in electrical engineering in 2004, and the Ph.D. degree in 2009, both from University of Malta, Msida, Malta. In January 2008, he was appointed Assistant Lecturer with the same department and is currently a Senior Lecturer. He has been in technical and organizational committees of several national and international conferences. In particular, he served as the General-Chair on the IEEE International Workshop on Biometrics and Forensics and as Technical Programme Co-Chair on the IEEE VISUAL COMMUNICATIONS AND IMAGE PROCESSING in 2014. He has been contributing as a Reviewer of several journals and conferences, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO AND TECHNOLOGY, and IEEE TRANSACTIONS ON MULTIMEDIA. On September 2013, he was appointed as the National Contact Point of the European Association of Biometrics.



Christian Galea (S'14) received the first degree in communications and computer engineering in 2012 and the Master's degree in telecommunications in 2014 both from the University of Malta, Msida, Malta. He is currently working toward the Ph.D. degree with the Faculty of Information and Communication Technology, University of Malta. His research interests include computer vision and image and video processing, especially applications related to objective image/video quality assessment, biometrics and automobiles.



Christine Guillemot (F'13) received the Ph.D. degree from ENST (Ecole Nationale Superieure des Telecommunications) Paris, Paris, France, and an Habilitation for Research Direction from the University of Rennes, Rennes, France. He is the Director of Research at INRIA. From 1985 to October 1997, she has been with France Telecom, where she has been involved in various projects in the area of image and video coding for TV, HDTV, and multimedia. From January 1990 to mid 1991, she has worked at Bellcore, NJ, USA, as a Visiting Scientist. Her research interests are signal and image processing, and in particular two dimensional (2-D) and 3-D image and video processing for various problems (compression, super-resolution, inpainting, classification). She has served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING (2000–2003, 2014–2016), for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (2004–2006), and for the IEEE TRANSACTIONS ON SIGNAL PROCESSING (2007–2009). She has served as a Senior Member of the Editorial Board of the IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING (2013–2015) and is currently a Senior Area Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING.

Image Reconstruction via Manifold Constrained Convolutional Sparse Coding for Image Sets

Linlin Yang, Ce Li, Jungong Han, Chen Chen, Qixiang Ye, Baochang Zhang, Xianbin Cao, and Wanquan Liu

***Abstract*—Convolution sparse coding (CSC) has attracted much attention recently due to its advantages in image reconstruction and enhancement. However, the coding process suffers from perturbations caused by variations of input samples, as the consistence of features from similar input samples are not well addressed in the existing literature. In this paper, we will tackle this feature consistence problem from a set of samples via a proposed manifold constrained convolutional sparse coding (MCSC) method. The core idea of MCSC is to use the intrinsic manifold (Laplacian) structure of the input data to regularize the traditional CSC such that the consistence between features extracted from input samples can be well preserved. To implement the proposed MCSC method efficiently, the alternating direction method of multipliers (ADMM) approach is employed, which can consistently integrate the underlying Laplacian constraints during the optimization process. With this regularized data structure constraint, the MCSC can achieve a much better solution which is robust to the variance of the input samples against overcomplete filters. We demonstrate the capacity of MCSC by providing the state-of-the-art results when applied it to the task of reconstructing light fields. Finally, we show that the proposed MCSC is a generic approach as it also achieves better results than the state-of-the-art approaches based on convolutional sparse coding in other image reconstruction tasks, such as face reconstruction, digit reconstruction, and image restoration.**

***Index Terms*—Convolutional sparse coding, image reconstruction, image deblurring, light field, manifold constrained convolutional sparse coding.**

Manuscript received December 30, 2016; revised May 15, 2017 and July 17, 2017; accepted August 16, 2017. Date of publication August 23, 2017; date of current version October 23, 2017. This work was supported in part by the Natural Science Foundation of China under Contracts 61672079, 61601466, and 61473086. The work of B. Zhang was supported in part by the Beijing Municipal Science and Technology Commission under Grant Z161100001616005. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Lu Fang. *(Linlin Yang and Ce Li contributed equally to this work.) (Corresponding author: Baochang Zhang.)*

L. Yang and B. Zhang are with the Department of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China (e-mail: mu4yang@qq.com; bczhang@buaa.edu.cn).

C. Li is with the China University of Mining and Technology, Beijing 100083, China (e-mail: 409719991@qq.com).

J. Han is with the School of Computing and Communications, Lancaster University, Lancaster LA1 4WA, U.K. (e-mail: jungonghan77@gmail.com).

Q. Ye is with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: qxye@ucas.edu.cn).

C. Chen is with the University of Central Florida, Orlando, FL 32816 USA (e-mail: chenchen870713@gmail.com).

X. Cao is with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China (e-mail: xbcao@buaa.edu.cn).

W. Liu is with the Curtin University, Perth, WA 6845, Australia (e-mail: wanquan@cs.curtin.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2743683

I. Introduction

With the increasing number of images and videos produced in our everyday life, batch data processing techniques which can deal with high-volume data containing significant illumination variation, blurring, and severe noise, are highly demanded. For instance, high-resolution light field imaging is a typical batch system applied on a set of images captured from large camera arrays or a hand-held camera moved over time around the scene [1], [2]. To handle the challenging problem caused by high sampling rate and huge-volume samples captured in various of external conditions (i.e., lighting variations), different kinds of techniques that can reconstruct dense light fields from fewer samples are developed based on various prior assumptions made about the data [2]. When processing batch images, researchers reach a consensus that the data structure as a common prior should be given much attention, which has been evidenced first in visualizing high-dimensional data in the low-dimension space [3], [4], and then some successful applications in the learning paradigm [5]–[7]. Especially, the data structure embedded into a constraint model could be used to simplify the learning process while improving the solution quality by the regularization technique, e.g., manifold regularization [5], which can be found in the latest learning algorithms [6], [7]. As detailed in [7], the constraints derived from the manifold structure in the input data can be used to improve the performance of convolution neural network (CNN). This is more important when one has to gain the robustness to variations of the input data. To this end, the methods better taking advantage of the implicit data-set structure can obtain higher performance.

In this paper, we revisit the convolution sparse coding (CSC) [8] by considering the data on a Laplacian manifold. The CSC or deconvolution, as one of fundamental tools in machine learning, has been demonstrated to have important applications in a broad range of computer vision problems including feature learning, particularly related to the pre-training of the CNN filters [9], image deionising and/or reconstruction [9]–[12] , super-resolution [13], trajectory reconstruction [14], automatic music transcription [15] and neuronal ensemble identification [16]. The conventional CSC method was improved by new coding or dictionary learning approaches in theory [17]–[19]. Based on the convolution operator and sparse coding, the CSC reconstructs the vectorized input images with vectorized sparse feature maps and convolutional kernels. There are some works addressing the deficiency of the CSC based on an alternating direction method of multipliers (ADMM) approach with the mixing matrix [8]. In [20], the extended CSC has been proposed by considering local features dependency for image reconstruction as a potential di-

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

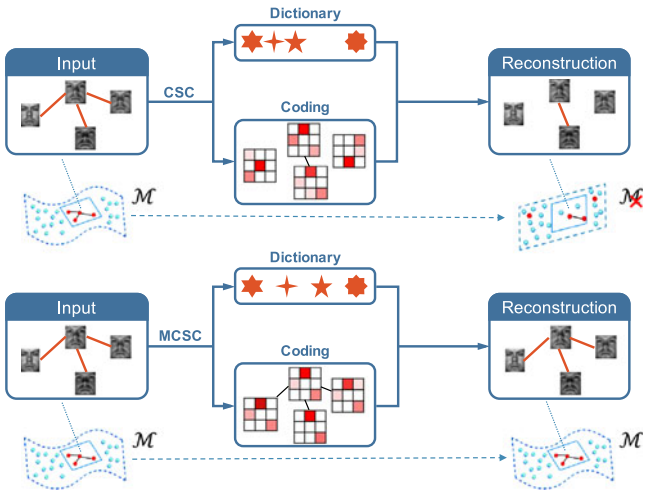


Fig. 1. Illustration of CSC and MCSC for image reconstruction. Reconstructed images by CSC that lacks structure constraint suffer from the perturbation caused by the noised or corrupted input samples. In contrast, the reconstructed images by MCSC with manifold constraint is more consistent with the input data structure. As a property of MCSC, the manifold structure of input data is preserved, leading to more reasonable reconstruction results.

rection of CSC, but the authors there only sketched out the idea with neither implementation details nor sufficient experimental results. Nevertheless, the data structure of input samples is not fully utilized in the existing CSC methods and the consistence of the features for similar input samples cannot be guaranteed. In other words, the coding results for similar input samples may vary dramatically because the neighborhood information in the data-set is ignored. With the over-complete filter set, the coding process without consistence constraints would suffer from the perturbations caused by the variations of input samples as they might be distorted or corrupted. As shown in Fig. 1, the neighborhood of input images in the reconstruction task is emphasized via a manifold structure constraint.

In this paper, we will consider the mutual dependence among observations, which would result in large variations in the reconstructed data once it is ignored and then this provides some new insights on the CSC with manifold constraints. In particular we have an intuition on the data-set structure that was neglected: the data lying on a specific manifold can be used to improve solutions for some applications. For such purpose, we propose a new scheme of the CSC as shown in Fig. 1 for image reconstruction. Being different from the traditional CSC without structure constraint, the coding process of the proposed manifold constrained convolutional sparse coding (MCSC) is imposed to be consistent with the data structure, while the learning algorithm is implemented with the regularization technique to solve the ill-posed problem. As an important property of the MCSC, the manifold structure can be preserved, whilst leading to a better reconstruction than the conventional CSC. Furthermore, based on recent advantages in ADMM, we can efficiently solve the proposed MCSC problem based on a new manifold constraint model. One should notice that there exist some previous works to incorporate the manifold structure into deep learning [7] or sparse coding [6], [21]. However, to the best of our knowledge, this is the first time that manifold constraint is introduced for the CSC. In summary, we have the following contributions in this paper.

1) The manifold structure considered as a constraint is embedded into the CSC and the proposed problem is theoretically investigated, by which the consistence of features for similar input samples can be preserved. Consequently, the reconstruction performance is improved as the implicit manifold structure is preserved.
2) The derived theoretical results show that imposing Laplacian manifold constraint on the input data is equivalent to imposing regularization terms on the coding and dictionary, which provides an insight for other similar methods [6], [21].
3) Extensive experimental results on image reconstruction, image de-blurring and light field reconstruction show a consistent improvement of performance with respect to the existing state-of-the-art algorithms based on CSC.

The rest of the paper is organized as follows. Section II introduces the related works, and Section III describes the details of the proposed method. Experiments and results are presented in Sections IV and V concludes the paper.

II. Related Work

Different from existing sparse coding algorithms for classification, the reconstruction and coding tasks that assume independent observations during the learning procedure, the conventional CSC explicitly models local interactions through the convolution operator, and the resulting optimization is considerably more complicated than traditional sparse coding, especially in light field reconstruction.

In [9]–[12], some hierarchical models that learn image decompositions via the CSC are proposed in feature learning. Moreover, an efficient feed-forward encoder that can predict sparse features was proposed for classification, which is particularly used to train the CNN filters [9], [22]. With the guarantee that neural activity is sparse in both space and time, Andilla *et al.* [16] proposed a sparse deconvolution approach in both space and time domain for calcium image sequences. Serrano *et al.* [23] verified the CSC on the high-speed video analysis task. The CSC also has attracted growing attention in processing different reconstruction tasks. Wohlberg *et al.* [17] proposed an extended CSC algorithm by using an augmented dictionary to incorporate the impulse noise images. Heide *et al.* [8] proposed a splitting-based ADMM approach with the mixing matrix to process the incomplete images and Serrano *et al.* [18] proposed a variant of CSC to recover high dynamic range images (HDRI), which achieved superior performances. In addition to the applications mentioned above, the CSC also has been utilized in super-resolution [13], trajectory reconstruction [14], automatic music transcription [15] and neuronal ensemble identification [16]. Meanwhile, based on recent advances in optimization [24], [25], varieties of efficient approaches to solve the CSC problems have been proposed, such as, the frequency domain method [26] and the splitting-based ADMM approach [8].

Recently, researchers have started to explore compressive light field acquisition with a single camera. The optical coding strategies include coded apertures [27], [28], coded

lenslets [27], and a combination of coded mask and aperture [29].The conventional methods require multiple images to be recorded and are not suitable for reconstructing dynamic scenes, even though they succeed in reducing the number of shots compared to their non-compressive counterparts. Marwah *et al.* [1] used sparse coding techniques to learn a dictionary of basis vectors for representing light fields, in which the dictionary is chosen so that training light fields may be represented as sparse vectors. The underlying assumption is that new light fields will have similar structure to those in their training data. Similarly, Shi *et al.* [2] proposed the continuous sparse structure in the Fourier domain for light field reconstruction.

Different from the existing works, we present here a new method that exploits Laplacian structure as constraints in the ADMM framework, in which we emphasize on the dependence among samples into the CSC. As a theoretical contribution, we show that the Laplacian manifold structure from data samples can be considered as regularization terms imposed on coding coefficients and dictionary. This makes our work significantly different from [21], where the Laplacian regularization on coding coefficients is used only for sparse coding. Instead, we focus on the convolution sparse coding with Laplacian regularization on both coding and dictionary learning in the ADMM framework. Our method is also different from [20], which is based on an assumption that Laplacian regularization is imposed on only a single image, whereas we focus on a batch of images.

III. Manifold Constrained CSC

In this section, we present how manifold can be incorporated into the CSC as a constraint and then construct the MCSC. We then solve the MCSC problem in the ADMM framework with theoretical analysis.

A. Laplacian Manifold for the CSC

As one of increasingly important tools in machine learning and computer vision, the CSC method can be used to learn the features subsequently used for classification and reconstruction tasks. Based on the convolution operator and sparse coding, the CSC is formulated as:

$$\begin{aligned} \underset{\mathbf{d},\mathbf{z}}{\text{minimize}} \quad & \frac{1}{2}\sum_{n=1}^{N}\left\|\mathbf{y}^n-\sum_{k=1}^{K}\mathbf{d}_k*\mathbf{z}_k^n\right\|_2^2+\beta\sum_{n=1}^{N}\sum_{k=1}^{K}\|\mathbf{z}_k^n\|_1 \\ \text{subject to} \quad & \|\mathbf{d}_k\|_2^2\le 1 \quad \forall k\in\{1,\dots,K\}, \end{aligned} \tag{P1}$$

where $\mathbf{y}^n \in R^{L_1}$ are vectorized input images, $\mathbf{z}_k^n \in R^{L_1}$ are vectorized sparse feature maps and $\mathbf{d}_k \in R^{L_2}$ are the vectorized convolutional kernels. N is the number of input images, β is the sparsity parameter, K is the number of the kernels and $*$ is the 2D convolution operator defined on the vectorized inputs. The formulation (P1) can be generally customized for different applications. In the above formulation, it is assumed that $\mathbf{y}^n \in R^{L_1}$ are independent. We assume that the input samples $\mathbf{y}^n \in R^{L_1}$ are lying on a manifold. To utilize the data structure prior, we first reformulate (P1) as:

$$\begin{aligned} \underset{\mathbf{d},\mathbf{z}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{y}-\mathbf{H}\mathbf{v}\|_F^2+\beta\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\|\mathbf{z}_k^n\|_1+\frac{1}{2}\|\mathbf{z}_k^n\|_2^2\right) \\ \text{subject to} \quad & \|\mathbf{d}_k\|_2^2\le 1 \quad \forall k\in\{1,\dots,K\} \\ & \mathbf{v^n}\in\mathcal{M}, \end{aligned} \tag{P2}$$

where $\mathbf{v}^n=\sum_{k=1}^{K}\mathbf{d}_k*\mathbf{z}_k^n$, $\mathbf{v}=[\mathbf{v}^1,\dots,\mathbf{v}^N]$, $\mathbf{y}=[\mathbf{y}^1,\dots,\mathbf{y}^N]$. $\mathcal{M}$ is a set of samples representing the underlying manifold. To encourage a grouping effect, where strongly correlated filters tend to be in or out of the model together, we choose elastic net constraints [30] instead of sparsity constraints. $\mathbf{H}$ is a diagonal matrix whose diagonal entries are either 1 or 0, keeping or killing the corresponding pixels for image reconstruction, or a measurement matrix containing the sheared mask code of multiple recorded sensors for light filed, or a blur matrix given by point spread function (PSF) and boundary condition for deblurring, which allows us to gain the robustness in the CSC framework for the complex situations, such as motion blur, out-of-focus blur and illumination variation. As a basic intuition, the manifold structure contained in the data set $\mathbf{y}^n \in R^{L_1}$ should be remained in the convolutional coding process. In particular $\mathbf{y}^n \in \mathcal{M}$ are the observations of degraded images, and $\mathbf{v}^n \in \mathcal{M}$ are the reconstructions of $\mathbf{y}^n$ that are thus subject to a manifold constraint. Again we expect the dependence among samples would help us obtain better coding results as well as better solution for the filter learning.

It should be noted that the manifold structure in the original input data is usually not accurate due to the fact that the input data may suffer from noise or corruption. In comparison, we expect the manifold structure becomes more accurate based on $\mathbf{v}$ as a better reconstruction after removing noise or corruption. Moreover, we prove that the manifold constraints in terms of $\mathbf{v}$ can be interpreted to be the meaningful regularizations in ADMM, which can improve the kernels orthogonally and features neighbor relationship. The proof details are presented in Section III-C.

In order to solve (P2) with structure preservation, $\mathbf{v}^n \in \mathcal{M}$ is the only unsolved part. We use the manifold in a constrained model that is different from the most popular approaches such as the Local Linear Embedding (LLE) [3] and the Laplacian with a focus on the data structure visualization. In contrast, the Laplacian Eigenmap [4] provides a computationally efficient approach to nonlinear dimensionality reduction that has locality-preserving properties with a natural connection to clustering. Here, we calculate the Laplacian structure of $\mathbf{y}$ in the input space that can be transferred to $\mathbf{v}$ via a specific constraint function. Technically, the Laplacian matrix $\mathbf{L}$ [4] is used to replace $\mathbf{v}^n \in \mathcal{M}$ in P2 and thus we can obtain the following MCSC formulation.

$$\begin{aligned} \underset{\mathbf{d},\mathbf{z}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{y}-\mathbf{H}\mathbf{v}\|_F^2+\frac{\gamma}{2}tr(\mathbf{v}\mathbf{L}\mathbf{v}^T) \\ & +\beta\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\|\mathbf{z}_k^n\|_1+\frac{1}{2}\|\mathbf{z}_k^n\|_2^2\right) \\ \text{subject to} \quad & \|\mathbf{d}_k\|_2^2\le 1 \quad \forall k\in\{1,\dots,K\}, \end{aligned} \tag{1}$$

where $\mathbf{L}$ is a Laplacian matrix with a parameter γ, which weights the contribution of the regularization term. The solution to (1) with a new manifold constraint function $\frac{\gamma}{2}tr(\mathbf{vLv}^T)$ will be investigated based on the ADMM framework in the next subsection. It is actually the only part where manifold plays a role in (P2). According to [5], the nonlinear structure as a resulting regularization term can avoid over-fitting and ensure that a problem is well-posed by penalizing complicated solutions.

B. Solution to (P2) Based on the ADMM

The main optimization tool we use in solving this convex problem (1) is ADMM, which is a variant of the standard Augmented Lagrangian Multiplier (ALM) method that uses partial updates (similar to the Gauss-Seidel method for solving linear equations). The ADMM recently has received much more attention because of its adaptability to several problems [24]. By iteratively solving a set of simple convex optimization subproblems, one can obtain the final solution of (P2). More details about ADMM can be found in [8], [24], [25], [31]. The ADMM can solve the sub-convex problems for two or more variables when one treats one variable each time while other variables are fixed. By exploiting the merits of the ADMM, we develop three update steps as described below: coding update, dictionary update and structure update. The coding update aims for more precise reconstruction coefficients, dictionary update for better representation and structure update for data manifold structure constraint.

First, the coding update is achieved by:

$$\begin{aligned}\underset{\mathbf{z}}{\text{minimize}} \quad & \frac{1}{2}\left\|\mathbf{y}-\mathbf{Hv}\right\|_F^2+\frac{\gamma}{2}tr(\mathbf{vLv}^T) \\ & +\beta\sum_{n=1}^{N}\sum_{k=1}^{K}\left(\left\|\mathbf{z}_k^n\right\|_1+\frac{1}{2}\left\|\mathbf{z}_k^n\right\|_2^2\right), \end{aligned} \tag{2}$$

where $\mathbf{v}^n=\mathbf{Dz}^n$, $\mathbf{D}=[\mathbf{D}_1,\ldots,\mathbf{D}_K]$ is a concatenation of Toeplitz matrix representing a convolution with the respective filter $\mathbf{d}_k$, $\mathbf{z}=[\mathbf{z}^1,\ldots,\mathbf{z}^N]$, $\mathbf{z}^n=[\mathbf{z}_1^{nT},\ldots,\mathbf{z}_k^{nT}]^T$.

Next, the dictionary update problem is defined as:

$$\begin{aligned}\underset{\mathbf{d}}{\text{minimize}} \quad & \frac{1}{2}\left\|\mathbf{y}-\mathbf{Hv}\right\|_F^2+\frac{\gamma}{2}tr(\mathbf{vLv}^T) \\ \text{subject to} \quad & \left\|\mathbf{d}_k\right\|_2^2\le 1 \quad \forall k\in\{1,\ldots,K\}, \end{aligned} \tag{3}$$

where $\mathbf{v}^n=\mathbf{Z}^n\mathbf{d}$, $\mathbf{Z}^n$ is a concatenation of Toeplitz matrices for the respective sparse codes $\mathbf{z}_k^n$, $\mathbf{Z}=[\mathbf{Z}_1,\ldots,\mathbf{Z}_N]$, $\mathbf{d}=[\mathbf{d}_1^T,\ldots,\mathbf{d}_K^T]^T$.

And finally, the structure update is shown as:

$$\mathbf{L}=(1-\eta)\mathbf{L}+\eta LE(\mathbf{Dz}), \tag{4}$$

where $\mathbf{L}$ is initialized equal to $LE(\mathbf{y})$, and $LE(\cdot)$ means normalized Laplacian matrix function defined in [4] and η represents update ratio.

Algorithm 1: ADMM for (P2).

1: Input image set $\mathbf{y}$ and parameters
2: Pre-compute $\mathbf{L}=LE(\mathbf{y}),\mathbf{H}$
3: Initialize $\mathbf{d},\mathbf{z}$ randomly
4: **repeat**
5: **Coding update**

$$\mathbf{z}=\underset{\mathbf{z}}{argmin}f_1(\mathbf{Dz})+f_2(\mathbf{Dz})+\sum_{n=1}^{N}\sum_{k=1}^{K}f_3(\mathbf{z}_k^n)$$

6: **Dictionary update**

$$\mathbf{d}=\underset{\mathbf{d}}{argmin}f_1(\mathbf{Zd})+f_2(\mathbf{Zd})+\sum_{k=1}^{K}f_4(\mathbf{d}_k)$$

7: **Structure update**

$$\mathbf{L}=(1-\eta)\mathbf{L}+\eta LE(\mathbf{Dz})$$

8: **until** The maximum number of iteration 50

To simplify our ADMM, we define four functions that are used to reformulate three mentioned update steps:

$$\begin{aligned} f_1(\mathbf{x})&=\frac{1}{2}\left\|\mathbf{y}-\mathbf{Hx}\right\|_2^2, & f_2(\mathbf{x})&=\frac{\gamma}{2}tr(\mathbf{xLx}^T), \\ f_3(\mathbf{x})&=\beta\left(\left\|\mathbf{x}\right\|_1+\frac{1}{2}\left\|\mathbf{x}\right\|_2^2\right), & f_4(\mathbf{x})&=ind_c(\mathbf{x}), \end{aligned} \tag{5}$$

where $ind_c(\cdot)$ represents the indicator function and $C=\{\mathbf{x}|\left\|\mathbf{Sx}\right\|_2^2\le 1\}$. The proximal operators [25] used to approximate the above four objectives can be derived as

$$\begin{aligned} \text{prox}_{\lambda f_1}(\mathbf{x})&=(\mathbf{I}+\lambda\mathbf{H^T}\mathbf{H})^{-1}(\mathbf{x}+\lambda\mathbf{H^T}\mathbf{y}) \\ \text{prox}_{\lambda f_2}(\mathbf{x})&=\mathbf{x}(\gamma\lambda\mathbf{L}+\mathbf{I})^{-1} \\ \text{prox}_{\lambda f_3}(\mathbf{x})&=\left(\frac{1}{1+\lambda}\right)\left(\max\left(1-\frac{\lambda}{|\mathbf{x}|},0\right)\odot\mathbf{x}\right) \\ \text{prox}_{\lambda f_4}(\mathbf{x})&=\begin{cases}\dfrac{\mathbf{Sx}}{\left\|\mathbf{Sx}\right\|_2^2} & \left\|\mathbf{Sx}\right\|_2^2\ge 1 \\ \mathbf{Sx} & \text{else}\end{cases}. \end{aligned}$$

In Algorithm 1, we show the iterative steps of the MCSC which can solve P2 efficiently.

C. Theoretic Analysis of the MCSC

In this subsection, we conduct theoretical analysis for the proposed MCSC in two theorems. In detail, Theorem 1 states that the Laplacian structure in input data $\mathbf{v}$ can approximately constrain the solution as that in feature space $\mathbf{z}$ (minimizing $tr(\mathbf{v}L\mathbf{v}^T)$ is approximated as $tr(\mathbf{z}L\mathbf{z}^T)$). Theorem 2 proves that in the dictionary update step the Laplacian structure constraint in input data is in fact equal to the constraint imposed on the dictionary ($l2$-norm). We describe two lemmas [32] and then prove our main theorems. The proofs of lemmas and theorems are shown in the Appendix.

Theorem 1: Given $\mathbf{D}$, and a positive constant α', we have:

$$c_1 tr(\mathbf{z}\mathbf{L}\mathbf{z}^T) \leq tr(\mathbf{v}\mathbf{L}\mathbf{v}^T) + \alpha' \sum_{n=1}^{N}\sum_{k=1}^{K} \|\mathbf{z}_k^n\|_2^2$$

$$\leq c_2 tr(\mathbf{z}\mathbf{L}\mathbf{z}^T) + \alpha' \sum_{n=1}^{N}\sum_{k=1}^{K} \|\mathbf{z}_k^n\|_2^2 ,$$

where c_1 and c_2 are both positive constants.

We observe in this theorem that as a property of the MCSC, minimizing $tr(\mathbf{v}\mathbf{L}\mathbf{v}^T)$ is approximated by minimizing $tr(\mathbf{z}\mathbf{L}\mathbf{z}^T)$. In other words the constraint on the data $\mathbf{v}$ can be transferred to the feature space $\mathbf{z}$. As mentioned above the Laplacian structure in the input data can constrain the solution as that in the feature space.

Theorem 2: Given $\mathbf{Z}$, and two positive constants α', β', we have:

$$c_3 \|\mathbf{d}\|_2^2 \leq tr(\mathbf{v}\mathbf{L}\mathbf{v}^T) + (\alpha' + \beta') \|\mathbf{d}\|_2^2 \leq c_4 \|\mathbf{d}\|_2^2 ,$$

where c_3 and c_4 are both positive constants. Similar to the observation in Theorem 1, minimizing $tr(\mathbf{v}\mathbf{L}\mathbf{v}^T)$ is approximated by minimizing $l2$-norm of $\mathbf{d}$. Again the constraint is transferred from one variable to another.

In summary, two theorems show that the Laplacian manifold structure of input data can be transferred as the regularization terms on optimized variables, coding coefficients or dictionary. Comparatively the nonlinear structure constraints in previous works are added on coding coefficients in a brute-force way [6], [21]. As another obvious advantage of MCSC, the manifold constraint on the data can be more explicitly represented by the optimized variables.

IV. Experiments

In this section, we use publicly available datasets to validate the effectiveness of the proposed approach on the reconstruction or restoration problems. The experiments are conducted on face (Yale B) [33], digits (USPS) [34], a real-application video for car [35], and the Chess [36]. The Yale B dataset consists of 2,414 frontal-face images from 38 individuals (about 64 images per subject) captured under various laboratory-controlled lighting conditions. The USPS dataset is an image dataset including 7291 training images and 2007 test images. The video for car consists of 900 97×79-sized images. The Chess [36] database consists of a 17×17 grid of viewpoints with 1400×800 image resolution, taken from the Stanford light field archive. The comparative methods are some state-of-the-art methods including patch based sparse coding (PSC) [1], [37], CSC [8], GSR [38] and MCSCf (see definition in Section IV-B). Following the experiments in [8], [37], one part of data is used for training the dictionary and the other is for testing in reconstruction tasks. To make manifold structure of the data more transparent, we first preprocessed the input images as follows:

- Scheme 1: Translating one and two pixels in each of the four directions, and rotating the images by one to three degrees.
- Scheme 2: Random subsampling of the data.
- Scheme 3: Adding motion blur with 0 mean and 0.01 deviations.

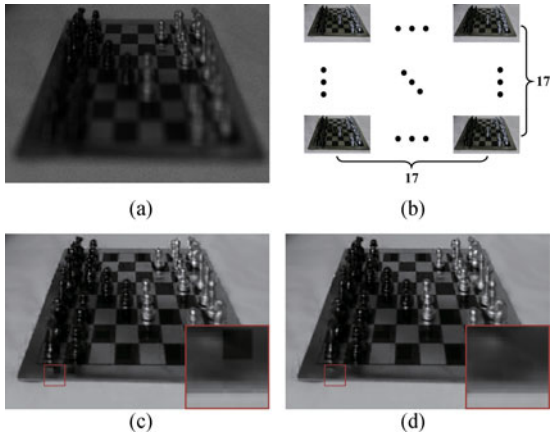


Fig. 2. (a) Coded 2D Projection. (b) Reconstructed 4D Light Field. (c) The reconstructed light field by CSC. (d) The reconstructed light field by MCSC.

TABLE I
Average PSNR (dB) for Chess and YaleB01 Database

Method	PSC	CSC	MCSCf	MCSC
Chess	29.79	30.96	–	**31.54**
YaleB01	27.78	28.19	28.40	**28.46**

To evaluate the quality of the reconstructed images, the signal-to-noise ratio (PSNR, unit:dB) is calculated. The higher PSNR is, the better visual quality is achieved for the reconstructed image. In our experiments, one part of data is used for training the dictionary and the other data is for testing in the reconstruction task. Three parameters are involved in the MCSC: the sparsity of coding β, the weight of manifold constraint γ and the manifold structure update ratio η. It should be noted that the parameters are set according to different applications which might be very different, aiming to deliver a good tradeoff between sparsity and data fit. Regarding to the computational cost, the MCSC can be solved as efficiently as that of the baseline algorithm [8]. For example, the running time for digit '0' in USPS are respectively 100 minutes and 96 minutes for the MCSC and CSC, on a PC with Intel i7 CPU and 16G RAM.

A. Light Field Reconstruction

We demonstrate an application of batch light field reconstruction on the widely used Chess database [36] for MCSC. The coded 2D projection and the reconstructed light fields are shown in Fig. 2. MCSC allows for higher-resolution light fields to be reconstructed than previously possible from a single image, where y is the vectorized observed sensor image, v is the vectorized light field, and H is a sparse measurement matrix containing the sheared coded projection operator on its diagonal, similar to [1]. In this case, there are 17×17 angular light field views stacked in v, the Laplacian matrix is initialized to a suitable default value and the resolution of sensor images are resized into 150×100 in our experiment. The recorded sensor image is stacked in y, and the corresponding measurement matrices are coded using random masks and stacked in H. The

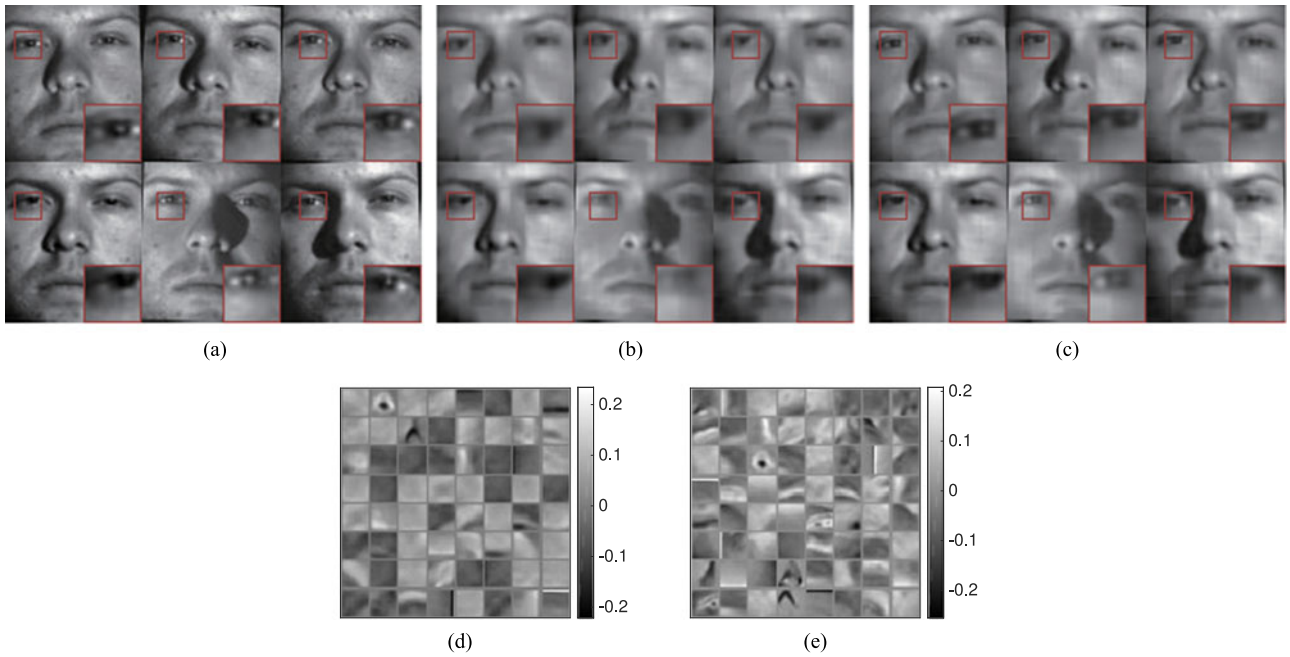


Fig. 3. (a) Original YaleB01 images. (b) Reconstruction by CSC. (c) Reconstruction by MCSC. (d) CSC dictionary. (e) MCSC dictionary.

interesting case of 4D light field reconstruction is addressed to validate the performance of MCSC, in comparison with the PSC [1] and CSC [8]. Following the baseline methods [1], [8], we train 64 8×8-sized filters, and then we set $\beta = 0.3$ in PSC and CSC, while $\beta = 0.01, \gamma = 0.1$ and $\eta = 0.01$ in MCSC. The results are shown in Fig. 2 and Table I. It can be observed that MCSC achieves better reconstruction quality (i.e., left-bottom of the chess board) than that of CSC, based on a compact dictionary due to the regularized manifold structure. The image can be well covered even under complex lighting conditions, occlusion and blurring.

B. Face Reconstruction

In this subsection, we evaluate the quality of face image reconstruction in terms of the average PSNR measure using the extended Yale Face Database B [33]. We process the images with scheme 1. The images are first resized to 96×84 and then normalized by dividing 255. We trained 64 15×15-sized filters, and compare these MCSC filters with the PSC [37] and CSC [8]. Besides, we also compare MCSC with the MCSCf filter. Specifically, MCSCf is defined as the CSC with the manifold constraint on the feature while MCSC is the CSC with the manifold constraint on both data and feature. We set $\beta = 0.3$ in the PSC and CSC, while $\beta = 0.3, \gamma = 0.1$ and $\eta = 0$ in MCSCf and MCSC. Six of the reconstructions and the dictionaries are shown in Fig. 3. Fig. 3(b) and (c) show the results for a set of six faces with different reconstruction methods, respectively. All left eyes (upper left) are amplified and compared in the bottom right corner. It can be observed that MCSC achieves much better visual quality (i.e., eyes) than that of CSC, because of the compactness of dictionary (Fig. 11) that contributes to better reconstruction results. In terms of the average PSNR values in Table I, MCSCf and MCSC improve CSC by 0.21 dB and 0.27 dB respectively. It is also obvious that the MCSC achieves a higher performance than other methods.

TABLE II
AVERAGE PSNR FOR DIFFERENT DIGIT DATASETS WITH 25 7×7-SIZED FILTERS

Digit	0	1	2	3	4
PSC	21.00	24.58	23.30	22.83	24.68
CSC	22.65	27.33	24.54	24.66	26.04
MCSCf	23.00	28.10	24.66	25.15	26.42
MCSC	**23.14**	**28.50**	**24.66**	**25.20**	**26.42**
Digit	5	6	7	8	9
PSC	23.17	23.28	24.29	23.33	24.27
CSC	24.56	24.79	25.59	24.87	26.08
MCSCf	24.74	25.06	26.14	24.99	26.35
MCSC	**24.95**	**25.23**	**26.40**	**25.14**	**26.42**

C. Digit Reconstruction

We perform handwritten digit reconstruction on the widely used the USPS database [34]. We experimented on digit '0' to '9' respectively with locally contrast normalization [8]. We adopt the parameters in [8] and set $\beta = 1$ in the PSC and CSC, while $\beta = 0.8$, $\gamma = 0.4$, $\eta = 0$ in MCSCf and MCSC. As baselines, the results of PSC [37] and CSC [8] are also listed in Table II. Moreover, the reconstructions for digit '9' have been shown in the amplified images in the bottom right corner of Fig. 4.[1] One can see that the proposed MCSC achieves a better performance than the compared methods.

To further validate the proposed method, we also conduct the experiment of digit recognition. In experiments, 200 training

[1]We show 100 out of the total reconstructions and dictionaries.

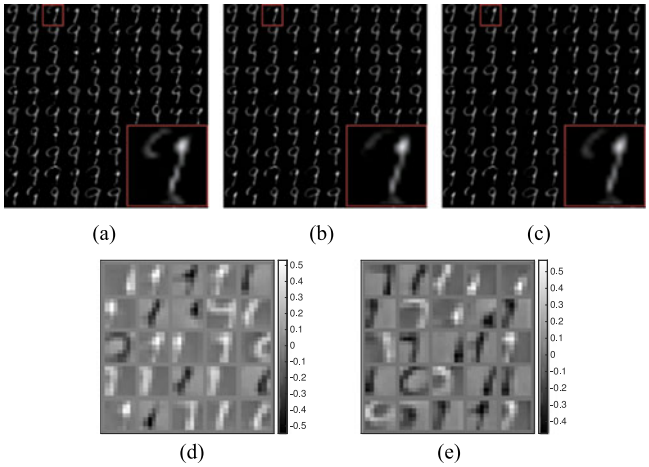


Fig. 4. (a) Original digit '9' images. (b) Reconstruction by CSC. (c) Reconstruction by MCSC. (d) CSC dictionary. (e) MCSC dictionary.

TABLE III
AVERAGE RECOGNITION ACCURACY (%) OF FOUR METHODS ON USPS DATABASE

Method	PSC	CSC	MCSCf	MCSC
Average Accuracy	97.11	97.50	97.95	**98.05**

images per class are randomly seletected for training and 2007 test images are selected for testing. 9 dictionaries with 7×7-sized filters are trained for each class, and the parameters are set in the same way as in [37]. We repreat this procedure 10 times. The average recognition accuracy shown in Table III presents the preliminary evidence of the capability of MCSC for pattern recognition task.

D. Image Restoration

Batch image restoration is to reconstruct the original image from degraded observations. As a typical application of the MCSC, one can apply it to different kind of restoration problems as long as appropriate H is given and also an offset term as a smoothness prior [8], [18]. Here, H is a diagnoal matrix or PSF, about the selection of H in our method is discussed for (P2). Two interesting cases of image restoration, i.e., inpainting and deblurring, are addressed to validate the performance of the proposed MCSC, in comparison with the CSC and group-based sparse representation (GSR) [38]. We optimally choose $\beta = 0.05$ in the GSR and CSC, while $\beta = 0.05$, $\gamma = 0.01$ and $\eta = 0.01$ in the MCSC. For inpainting, we process the digit '0' in USPS dataset with scheme 2. Inpainting restoration results of '0' (100 out of the total) are shown in the bottom right of Fig. 5 and Table IV. To further validate the effectiveness of the structure constraints, we test the parameters $\gamma = 0.01$, $\gamma = 0.5$, and $\gamma = 2$, for image inpainting as shown in Fig. 5(d)–(f), respectively. Obviously, the parameter γ influences the completeness of the digit '0', which again validates the effectiveness of the manifold constraint in the MCSC. For deblurring, we use the video car selected from a public dataset [35] that is further processed with scheme 3. The deblurring restoration results (2 out of 900) are compared in Fig. 6 and Table IV. Again, the MCSC achieves the best reconstruction results, due to the regularized manifold structure imposed on the CSC.

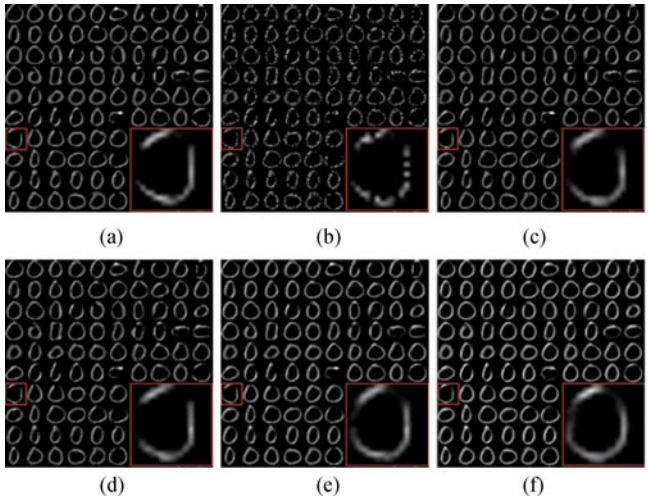


Fig. 5. Comparison of different inpainting methods for 75% subsampling. (a) Original digit '0' images. (b) Subsampled images. (c) The inpainted images by CSC. (d) The inpainted images by MCSC ($\gamma = 0.01$, $\eta = 0.01$). (e) The inpainted images by MCSC ($\gamma = 0.5$, $\eta = 0.01$). (f) The inpainted images by MCSC ($\gamma = 2$, $\eta = 0.01$).

TABLE IV
AVERAGE PSNR OF INPAINTING METHODS FOR DIFFERENT SUBSAMPLING RATES AND DEBLURRING METHODS

Method	Subsample	GSR	CSC	MCSC
Inpainting for 75%	22.31	34.64	35.16	35.44
Inpainting for 50%	19.35	29.31	30.19	30.59
Method	Motion blur	GSR	CSC	MCSC
Deblurring	23.37	33.89	34.13	34.24

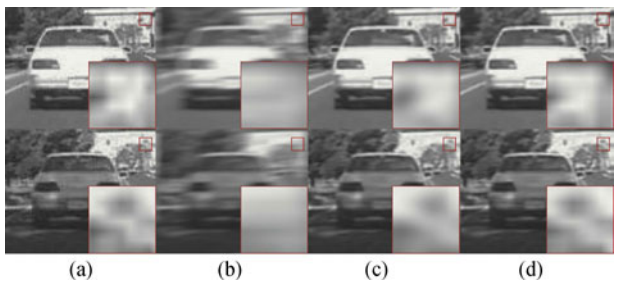


Fig. 6. Comparison of different deblurring methods. (a) Original car images. (b) Motion blur images. (c) The deblurred images by CSC. (d) The deblurred images by MCSC.

E. Discussion of the MCSC

As shown in the above experiments, MCSC achieves a higher performance than the well-known baseline CSC. This section further discusses how a higher performance has been achieved. We illustrate the performance of MCSC and CSC

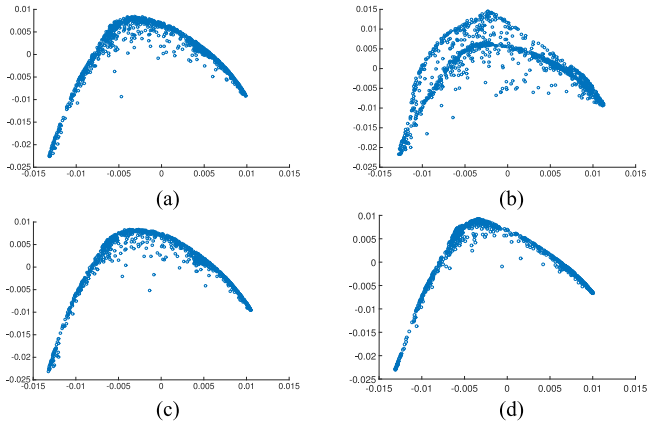


Fig. 7. Comparison of the manifold of '0' samples. (a) Manifold of the original input data. (b) Manifold of CSC. (c) Manifold of MCSC ($\gamma = 0.1, \eta = 0$). (d) Manifold of MCSC ($\gamma = 5, \eta = 0.5$).

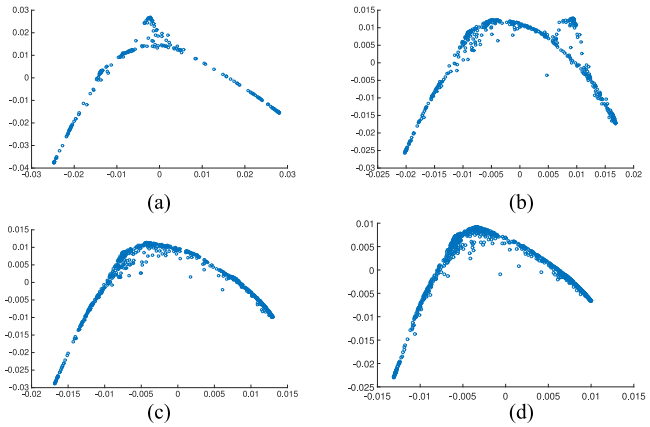


Fig. 8. Manifold of MCSC with different numbers of '0' samples. (a) 250 samples, (b) 500 samples, (c) 750 samples, (d) 1000 samples.

based on the reconstruction error, which is an important metric to evaluate for evaluating the quality of image reconstruction. A smaller value denotes less information loss. But when we process batch images, the reconstruction error can not be used to valuate the quality of data structure. Even for a same reconstruction error, manifold structures may vary a lot. Differently we focus on both reconstruction error and structure illustration. Imposing the manifold constraint is expected to better preserve the data structure and decrease the reconstruction loss.

Manifold Structure: In Fig. 7, we plot manifold structures of the input data, CSC and MCSC reconstructions on digit '0' in USPS. The dataset has more than 1000 images, based on which we estimate the structures using the Laplacian Eigenmaps [4] with the same iterations. Obviously, MCSC ($\gamma = 0.1, \eta = 0$) achieves more similar manifold structures as the input manifold than CSC, in comparison with Fig. 7(b), Fig. 7(c) is much more similar to Fig. 7(a). The result shown in Fig. 7(b) just gets worse reconstruction actually. In addition, we plot the manifold of MCSC with $\gamma = 5$ and $\eta = 0.5$ in Fig. 7(d), as expected that larger γ and η can improve the smoothness of the structure. Besides γ and η, the number of samples also plays a crucial role in the structure calculation, and the results in Fig. 8 confirm that larger sample sizes result in better structures. We further select 2000 pairwise of digit '6' and evaluate the normalized correlation between the input image and its reconstruction in Fig. 9. One can see that MCSC shows a better linear fitting of $y = x$ while CSC has a bigger bias and variance.

Reconstruction Error: With respect to the reconstruction error, we calculate it in the YaleB01 dataset. In Fig. 10, the MCSC demonstrates the capability of decreasing the reconstruction loss. In addition to the manifold and reconstruction error analysis mentioned above, we also evaluate the performance on dictionary compactness and feature similarity. As theorem 1 and 2 showed that the imposing manifold constraint should have the impact on compacting dictionaries and preserving feature similarity.

Dictionary Orthogonality: To evaluate the compactness of the learned convolutional filters, we define the normalized

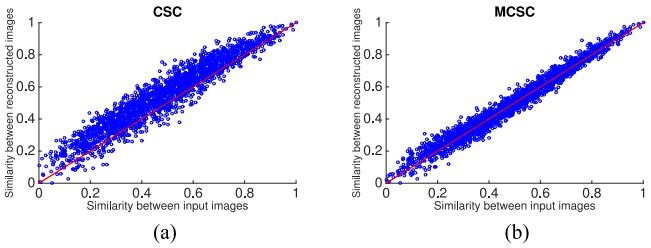


Fig. 9. (a) The similarity correspondence between the inputs and the CSC reconstructions. (b)The similarity correspondence between the inputs and the MCSC reconstructions.

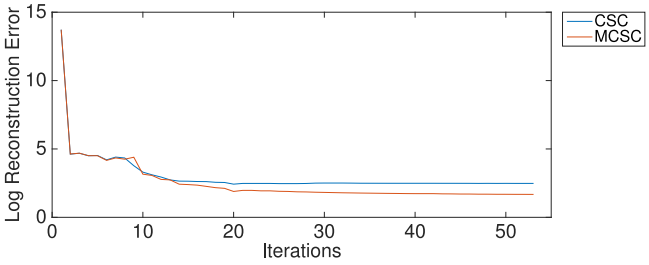


Fig. 10. Reconstruction error analysis.

dictionary orthogonality function for $\mathbf{dic}$ as

$$\text{orth}(\mathbf{dic}) = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \mathbf{dic}_i^T \mathbf{dic}_j,$$

where $\mathbf{dic}_i$, the ith column of the $\mathbf{dic}$, represents a vectorized and normalized word. We calculate the normalized orthogonality in digit '6' with different kernel numbers. As shown in Fig. 11, the MCSC is more compact than that of CSC in terms of the dictionary orthogonality in most cases.

Feature Similarity: For feature similarity analysis, we plot the results of CSC and MCSC on a pair of images (the same class) sampled from YaleB data set in Fig. 12. It can be seen that the responses using the MCSC are much more similar with a higher similarity score, but the responses using the CSC vary a lot. Selecting 2000 pairwise handwritten numbers from digit '6' in

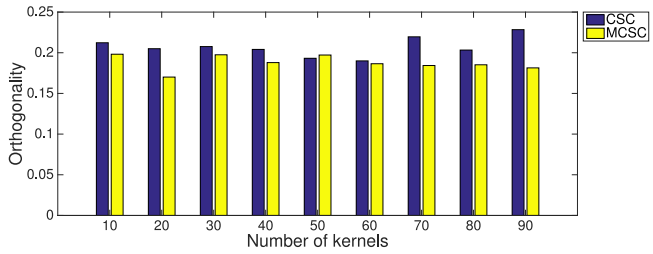


Fig. 11. Dictionary normalized orthogonal analysis with different kernel numbers of digit '6'.

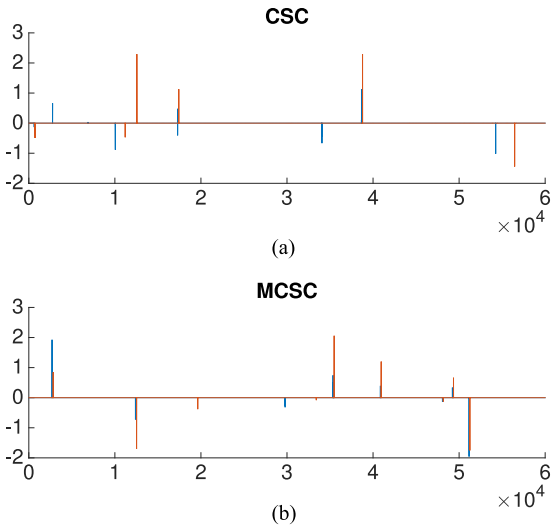


Fig. 12. (a) CSC feature maps (similarity = 0.25). (b) MCSC feature maps (similarity = 0.37). Noted that the similarity for the original image pair is 0.75.

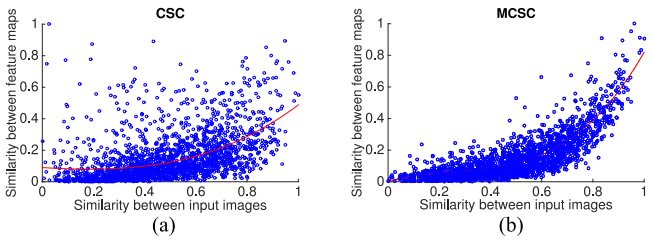


Fig. 13. (a) The similarity correspondence between the inputs and the CSC feature maps. (b) The similarity correspondence between the inputs and the MCSC feature maps.

USPS dataset, we calculate the relationship between their input and feature maps using the normalized correlation, and plot their similarity correspondence in Fig. 13. The MCSC similarity in Fig. 13(b) appears better fitting to $y = x$ than CSC in Fig. 13(a).

V. Conclusion

In this paper, we propose a novel framework of convolution sparse coding with Laplacian manifold constraint. The new insight of MCSC investigates mutual dependence among the observed and the reconstructed instances. Some theoretical results are obtained to explain the existing Laplacian regularization technique in sparse coding, and the reconstruction experimental results, especially in light field, show that the proposed algorithm performs favorably better against the state-of-the-art methods based on convolutional sparse coding. In future, we will investigate the new idea in this paper for tasks of feature extractions [39], classifications [40] and tracking [41], [42].

Appendix

Lemma 1: If $\mathbf{A} \in R^{n\times n}$ and $\mathbf{B} \in R^{n\times n}$ are positive semidefinite, then

$$tr(\mathbf{AB}) \leq tr(\mathbf{A})tr(\mathbf{B}).$$

Lemma 2: If $\mathbf{A} \in R^{n\times n}$ is positive definite and $\mathbf{B} \in R^{n\times n}$ is positive semidefinite, then

$$tr(\mathbf{AB}) \geq \frac{tr(\mathbf{B})}{tr(\mathbf{A}^{-1})}.$$

Proof of Theorem 1: In the coding update, $\mathbf{D}$ is given and $\mathbf{L}$ is positive semidefinite, we have: $tr(\mathbf{vLv}^T) = tr(\mathbf{DzLz}^T\mathbf{D}^T) = tr(\mathbf{D}^T\mathbf{DzLz}^T)$, where $\mathbf{D}^T\mathbf{D}$ is positive semidefinite, and $\mathbf{zLz}^T$ positive semidefinite. We then have: ■

$$tr(\mathbf{vLv}^T) + \alpha' \sum_{n=1}^{N}\sum_{k=1}^{K} \|\mathbf{z}_k^n\|_2^2$$

$$\geq tr(\mathbf{D}^T\mathbf{DzLz}^T) + \frac{\alpha'}{tr(\mathbf{L})}tr(\mathbf{zLz}^T) \quad \geq c_1 tr(\mathbf{z}^T\mathbf{Lz})$$

and

$$tr(\mathbf{vLv}^T) \leq tr(\mathbf{D}^T\mathbf{D})tr(\mathbf{zLz}^T) = c_2 tr(\mathbf{zLz}^T),$$

where $c_1 = 1/tr((\mathbf{D}^T\mathbf{D} + \frac{\alpha'}{tr(\mathbf{L})}\mathbf{I})^{-1})$, $c_2 = tr(\mathbf{D}^T\mathbf{D})$.

Proof of Theorem 2: In the dictionary update step, $\mathbf{Z}$ is given and $\mathbf{L}$ is positive semidefinite, and we have: $tr(\mathbf{vLv}^T) = tr(\mathbf{Z}\widetilde{\mathbf{D}}\mathbf{L}\widetilde{\mathbf{D}}^T\mathbf{Z}^T) = tr(\mathbf{Z}^T\mathbf{Z}\widetilde{\mathbf{D}}\mathbf{L}\widetilde{\mathbf{D}}^T)$, where $\widetilde{\mathbf{D}}$ is a diagonal matrix, whose diagonal entries are denoted by $\mathbf{d}$, and the zero vector has the same size of vector $\mathbf{d}$. $\|\widetilde{\mathbf{D}}\|_F^2 = N\|\mathbf{d}\|_2^2$. As $\mathbf{Z}^T\mathbf{Z}$ and $\widetilde{\mathbf{D}}\mathbf{L}\widetilde{\mathbf{D}}^T$ are both positive semidefinite, we have: ■

$$tr(\mathbf{vLv}^T) + (\alpha' + \beta')\|\mathbf{d}\|_2^2$$

$$\geq tr(\mathbf{vLv}^T) + \frac{\alpha'}{Ntr(\mathbf{L})}tr(\widetilde{\mathbf{D}}\mathbf{L}\widetilde{\mathbf{D}}^T) + \frac{\beta'}{N}\|\widetilde{\mathbf{D}}\|_F^2$$

$$\geq \frac{tr(\widetilde{\mathbf{D}}\mathbf{L}\widetilde{\mathbf{D}}^T)}{tr((\mathbf{Z}^T\mathbf{Z} + \frac{\alpha'}{Ntr(\mathbf{L})}\mathbf{I})^{-1})} + \frac{\beta'}{N}tr(\widetilde{\mathbf{D}}^T\widetilde{\mathbf{D}}) \geq c_3\|\mathbf{d}\|_2^2,$$

and

$$tr(\mathbf{vLv}^T) + (\alpha' + \beta')\|\mathbf{d}\|_2^2$$

$$\leq tr(\mathbf{Z}^T\mathbf{Z})tr(\mathbf{L})tr(\widetilde{\mathbf{D}}\widetilde{\mathbf{D}}^T) + (\alpha' + \beta')\|\mathbf{d}\|_2^2 = c_4\|\mathbf{d}\|_2^2,$$

where $c_3 = 1/tr((\frac{N\mathbf{L}}{tr((\mathbf{Z}^T\mathbf{Z} + \frac{\alpha'}{Ntr(\mathbf{L})}\mathbf{I})^{-1})} + \beta'\mathbf{I})^{-1})$, $c_4 = Ntr(\mathbf{Z}^T\mathbf{Z})tr(\mathbf{L}) + \alpha' + \beta'$.

REFERENCES

[1] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections," *ACM Trans. Graph.*, vol. 32, no. 4, 2013, Art. no. 46.

[2] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous fourier domain," *ACM Trans. Graph.*, vol. 34, no. 1, 2014, Art. no. 12.

[3] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2001, vol. 14, pp. 585–591.

[5] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, no. Nov, pp. 2399–2434, 2006.

[6] B. Zhang, A. Perina, V. Murino, and A. Del Bue, "Sparse representation classification with manifold constraints transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4557–4565.

[7] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou, "Multi-manifold deep metric learning for image set classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1137–1145.

[8] F. Heide, W. Heidrich, and G. Wetzstein, "Fast and flexible convolutional sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog*. IEEE, 2015, pp. 5135–5143.

[9] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3626–3633.

[10] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2018–2025.

[11] Y. Zhou, H. Chang, K. Barner, P. Spellman, and B. Parvin, "Classification of histology sections via multispectral convolutional sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3081–3088.

[12] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2528–2535.

[13] S. Gu *et al.*, "Convolutional sparse coding for image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1823–1831.

[14] Y. Zhu and S. Lucey, "Convolutional sparse coding for trajectory reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 529–540, Mar. 2015.

[15] A. Cogliati, Z. Duan, and B. Wohlberg, "Piano music transcription with fast convolutional sparse coding," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2015, pp. 1–6.

[16] F. D. Andilla and F. A. Hamprecht, "Sparse space-time deconvolution for calcium image analysis," *Adv. Neural Inf. Process. Syst.*, 2014, pp. 64–72.

[17] B. Wohlberg, "Convolutional sparse representations as an image model for impulse noise restoration," in *Proc. IEEE Image, Video, Multidimensional Signal Process. Workshop*, 2016, pp. 1–5.

[18] A. Serrano, F. Heide, D. Gutierrez, G. Wetzstein, and B. Masia, "Convolutional sparse coding for high dynamic range imaging," *Comput. Graph. Forum*, vol. 35, pp. 153–163, 2016.

[19] F. Heide, L. Xiao, A. Kolb, M. B. Hullin, and W. Heidrich, "Imaging in scattering media using correlation image sensors and sparse convolutional coding," *Opt. Express*, vol. 22, no. 21, pp. 26338–26350, 2014.

[20] X. Luo and B. Wohlberg, "Convolutional laplacian sparse coding," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation*, 2016, pp. 133–136.

[21] S. Gao, I. W.-H. Tsang, L.-T. Chia, and P. Zhao, "Local features are not lonely–Laplacian sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3555–3561.

[22] X. Wang, R. B. Li, and J. Currey, "Leveraging 2d and 3d cues for fine-grained object classification," in *Proc. IEEE Int. Conf. Image Process.*, pp. 1354–1358, 2016.

[23] A. Serrano, E. Garces, D. Gutierrez, and B. Masia, "Convolutional sparse coding for capturing high-speed video content," *Comput. Graph. Forum*, 2017. [Online]. Available: http://webdiis.unizar.es/∼aserrano/projects/CSC-Video.html

[24] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[25] N. Parikh and S. P. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014.

[26] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 391–398.

[27] A. Ashok and M. A. Neifeld, "Compressive light field imaging," *Proc. SPIE*, vol. 7690, 2010, Art. no. 76900Q.

[28] S. D. Babacan, R. Ansorge, M. Luessi, P. R. Mataran, R. Molina, and A. K. Katsaggelos, "Compressive light field sensing," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4746–4757, Dec. 2012.

[29] Z. Xu and E. Y. Lam, "A high-resolution lightfield camera with dual-mask design," *Proc. SPIE*, vol. 8500, 2012, Art. no. 85000U.

[30] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., Ser. B (Statist. Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.

[31] L. Yang, C. Chen, B. Zhang, and J. Han, "Adaptive multi-class correlation filters," in *Proc. Pacific-Rim Conf. Multimedia*, 2016, pp. 680–688.

[32] R. Patel and M. Toda, "Trace inequalities involving Hermitian matrices," *Linear Algebr. Appl.*, vol. 23, pp. 13–20, 1979.

[33] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[34] J. J. Hull, "A database for handwritten text recognition research," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, May 1994.

[35] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.

[36] A. Adams, "Stanford light field archive," Comput. Graph. Lab., Stanford Univ., Stanford, CA, USA, 2008. [Online]. Available: http://lightfield.stanford.edu

[37] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, 2010.

[38] J. Zhang, D. Zhao, and W. Gao, "Group-based sparse representation for image restoration," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3336–3351, Aug. 2014.

[39] B. Zhang, A. Perina, Z. Li, V. Murino, J. Liu, and R. Ji, "Bounding multiple guaasian uncertainty with application to object tracking," *Int. J. Comput. Vis.*, vol. 118, no. 3, pp. 364–379, 2016.

[40] B. Zhang, Y. Yang, C. Chen, L. Yang, J. Han, and L. Shao, "Action recognition using 3d histograms of texture and a multi-class boosting classifier," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4648–4660, Oct. 2017.

[41] B. Zhang *et al.*, "Output constraint transfer for kernelized correlation filter in tracking," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 4, pp. 693–703, Apr. 2017.

[42] B. Zhang, Z. Li, A. Perina, A. D. Bue, V. Murino, and J. Liu, "Adaptive local movement modeling for robust object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 7, pp. 1515–1526, Jul. 2017.

Authors' photographs and biographies not available at the time of publication.

Accelerated Shearlet-Domain Light Field Reconstruction

Suren Vagharshakyan, *Member, IEEE*, Robert Bregovic, *Member, IEEE*, and Atanas Gotchev, *Member, IEEE*

***Abstract*—We consider the problem of reconstructing densely sampled light field (DSLF) from sparse camera views. In our previous work, the DSLF has been reconstructed by processing epipolar-plane images (EPI) employing sparse regularization in shearlet transform domain. With the aim to avoid redundant processing and reduce the overall reconstruction time, in this paper, we propose algorithm modifications in three directions. First, we modify the basic algorithm by offering a faster and more stable iterative procedure. Second, we elaborate on the proper use of color redundancy by studying the effect of reconstruction of an average intensity channel and its use as a guiding mode for colorizing the three color channels. Third, we explore similarities between EPIs by their grouping and joint processing or by effective decorrelation to get an initial estimate for the basic iterative procedure. We are specifically interested in GPU-based computations allowing an efficient implementation of the shearlet transform. We quantify our three main approaches to accelerated processing over a wide collection of horizontal as well as full-parallax datasets.**

***Index Terms*—Light field reconstruction, graphics processing units, densely sampled light field.**

I. Introduction

3D VISUAL scenes are completely represented by the light field they emanate. Given that the light field is a continuous function, its capture and consequent reconstruction is an important task, especially for visualization applications, which require multiple perspective views (e.g. super multiview displays [1]) or dense parallax (e.g. digitally printed holograms [2]). Many other light field image processing applications, such as depth estimation, compression, synthetic aperture imaging would benefit from accurately reconstructed light field [3]. A typical way of capturing light fields from real world scenes is to use a set of identical parallel cameras which are uniformly positioned on a plane. In order to support continuous parallax, such capturing setup requires that the cameras are densely positioned [4]. To overcome the demand for synchronously controlled high amount of cameras, the approach is to use a coarse set of cameras and to devise a consecutive light field reconstruction method, which can deliver densely sampled views from the coarse set of captured ones.

The approaches for reconstructing dense intermediate views from a given sparse set of views can be categorized into two categories. First, those are methods aimed at extracting geometry information about the scene in the of form of high quality depth maps collocated with the given input images, which can be used for depth-based view rendering [5] or unstructured lumigraph rendering [6]. Such methods utilize correspondences between images, found by block matching [7], and employ some global optimization of cost functions usually formed by data and smoothness terms [8], [9]. Apart from having problems with occlusions, when using sparse views, these methods result in over-smoothed depth estimates, and for finding finest details aligned with object boundaries they still need relatively densely positioned cameras [10]. Second category includes methods operating directly on the light field and aimed at employing some sparsity priors for this data. For example, the work [11] exploits sparse representation of full parallax 4D light field in continuous Fourier domain using a small number of 1D viewpoint trajectories.

For both categories, reconstructing a dense set of images is a computationally demanding problem. Global optimization methods aimed at obtaining multiple high quality depth maps do not scale well with the number of images and their resolution. The method in [10] targeted processing of 50 views with overall of 21-megapixels and accounted of about 50 mins of processing time. The runtime of the method in [11] ranged from 2 to 3 hours using a cluster of machines.

Previously, we have proposed a method for light field reconstruction which utilizes sparsification in shearlet transform domain [12], [13]. The method reconstructs DSLF from an undersampled light field captured by a small number of wide-baseline cameras. It demonstrated superior performance while compared with Motion Picture Experts Group's depth estimation reference software (DERS) [14] and view synthesis reference software (VSRS) [15], and with the state of the art in depth-from-stereo scene geometry reconstruction [16]. The method handles both horizontal and full-parallax capture settings and is highly successful when reconstructing non-Lambertian scenes formed by semi-transparent objects [13]. In this article, we further develop the method by proposing computational acceleration approaches based on inherent similarities in the assumed data representation and further algorithm tuning. Our aim is to decrease the necessary computational time while keeping or even increasing the reconstruction quality for a large set of test data.

Manuscript received December 30, 2016; revised May 28, 2017; accepted July 14, 2017. Date of publication August 11, 2017; date of current version October 23, 2017. The guest editor coordinating the review of this paper and approving it for publication was Prof. Yebin Liu. *(Corresponding author: Suren Vagharshakyan.)*

The authors are with the Department of Signal Processing, Tampere University of Technology, Tampere 33101, Finland (e-mail: suren.vagharshakyan@tut.fi; robert.bregovic@tut.fi; atanas.gotchev@tut.fi).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2738617

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

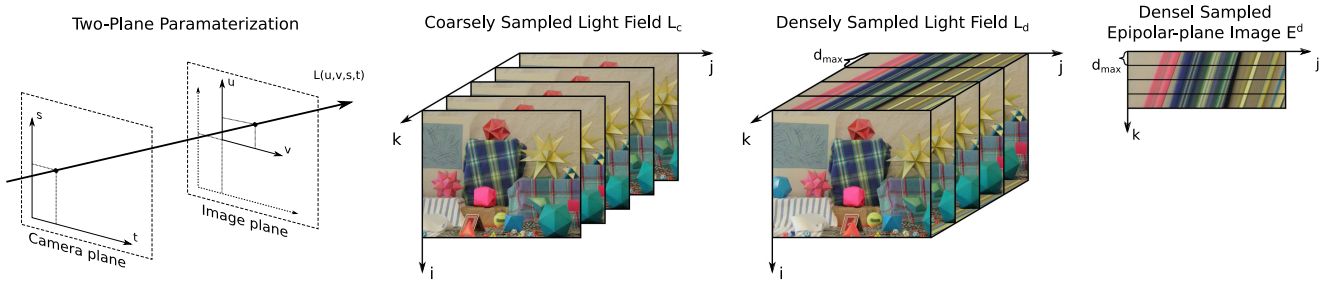


Fig. 1. Light field two-plane parameterization and corresponding discrete coarsely and densely sampled light field parameterizations.

The article is structured as follows: the light field parameterization and a summary of light field reconstruction algorithm from [13] are presented in Section II. Different acceleration approaches are proposed in Section III. Computing and evaluation setup, algorithm implementation, experimental results and discussions are presented in Section IV.

II. Reconstruction of Densely Sampled Light Field

4D light field is parameterized by the so-called two-plane parameterization $L(u, v, s, t)$ (Fig. 1), where (s, t) and (u, v) correspond to the camera plane and the image plane respectively [17]. This parameterization allows to conveniently describe and denote both the set of images (views) captured by a multi-camera setup and the required dense set of images fully representing the 3D scene of interest. For the sake of simpler notations and easier illustrations, hereafter we consider the case of horizontal parallax and explain the generalisations to full parallax when needed. Fixing the camera motion to horizontal direction only implies that the parameter $s = s_0$ corresponding to the vertical parallax can be omitted.

$$L(u, v, t) = L(u, v, s_0, t).$$

We aim at reconstructing continuous light field $L(u, s, t)$ from sampled views of 3D scenes. For such scenes and corresponding light fields, we define the densely sampled light field (DSLF), denoted by $L^d(i, j, k)$, as the light field having a maximal disparity between adjacent views less than 1 pixel. The k-th captured image $I^d_k(i, j) = L^d(i, j, k)$ corresponds to an image of the L continuous light field sampled at $t_k = k\Delta t_d$. The necessary parameter Δt_d for capturing DSLF can be calculated based on camera intrinsic parameters and specifying the minimum depth of the scene from the camera (capturing) plane. The continuous light field can be reconstructed from DSLF by linear interpolation [4]. Therefore, DSLF is a convenient representation and is in the core of many LF-based algorithms, such as refocusing, free view-point rendering and smooth-parallax visualization [3]. However, a direct capture of views providing one-pixel disparity is impractical.

In our previous work we have presented a method for DSLF reconstruction from coarsely sampled views [12]. A coarsely sampled light field is assumed to be a decimated version of DSLF, where the decimation factor is denoted by $d_{\max}$

$$L^c(i, j, k) = L^d(i, j, d_{\max}k).$$

subsequently, the maximal disparity between adjacent coarsely sampled views $I^c_k(i, j) = L^c(i, j, k)$ is no more than $d_{\max}$. In our previous work we presented an iterative algorithm which reconstructs L^d from L^c for $d_{\max} \leq 32$. The algorithm works in EPI domain. More specifically, DSLF L^d is reconstructed by reconstructing every densely sampled epipolar-plane image (DSEPI) defined as

$$E^d_i(k, j) = L^d(i, j, k)$$

from the given decimated samples E^c_i such that $E^c_i(k, j) = E^d_i(d_{\max}k, j)$. An example of coarsely-sampled and densely-sampled light fields is given in Fig. 1. Note how rows with step size $d_{\max}$ form E^c out of E^d. The specific value of $d_{\max} \leq 32$ is related with image resolution and has been selected for practical reasons. The method can be applied for higher disparity ranges too, however this would impose processing images with higher resolution, which in turn would significantly increase the required amount of memory [13]. Therefore, in this article we consider the limit case of $d_{\max} \leq 32$.

Below, we summarize the algorithm for DSEPI reconstruction [13]. To simplify the notations, we denote the unknown DSEPI matrix by $f \in \mathbb{R}^{N\times N}$. The decimated EPI $g \in \mathbb{R}^{N\times N}$ has the same dimension and contains sensed values at each $d_{\max}$-th row while the other rows are set to 0. The relation between the two EPIs is formalized by setting a binary measurement matrix $M \in \mathbb{R}^{N\times N}$ which has zero values elsewhere than $M(kd_{\max}, j) = 1$. Then, $g = M \odot f$, where $\odot$ is element-wise matrix multiplication. The direct and inverse shearlet transforms are denoted by $S : \mathbb{R}^{N\times N} \to \mathbb{R}^{\eta\times N\times N}$ and $S^* : \mathbb{R}^{\eta\times N\times N} \to \mathbb{R}^{N\times N}$, respectively, where η is the number of all shears in all scales of the shearlet transform. More details about the shearlet transform construction can be found in [13]. The reconstruction of unknown rows of the matrix g is formulated under the prior condition for having sparse solution in the shearlet domain, i.e.

$$\min_{f\in\mathbb{R}^{N\times N}} \|S(f)\|_0 \text{ , subject to } g = M \odot f$$

which can be efficiently solved through the following iterative thresholding algorithm [13]:

$$f_{n+1} = S^*(T_{\lambda_n}(S(f_n + \alpha_n(g - M \odot f_n)))), \qquad (1)$$

where the acceleration parameter α_n is chosen as follows

$$\alpha_n = \frac{\|\beta_n\|_2^2}{\|M \odot S^*(\beta_n)\|_2^2},$$
$$\beta_n = S_{\Gamma_n}(y - M \odot f_n), \Gamma_n = \text{supp}(f_n), \quad (2)$$

and $(T_\lambda f)(k) = \begin{cases} f(k), & |x(k)| \geq \lambda \\ 0, & |x(k)| < \lambda \end{cases}$ is a hard thresholding operator. The initial value can be set to $f_0 = S_0^*(S_0(g))$, where S_0 and S_0^* are direct and inverse transform using only low-pass element in the shearlet transform. The thresholding parameter λ_n is set to decrease with the iteration number n. In our case we apply a linear decrease from $\lambda_{\max}$ to $\lambda_{\min}$, for L iterations such as $n = 0, \ldots, L$.

It is important to mention that one has to set a few parameters while running the algorithm. These are the number of iterations L which directly influences the computational time; the initial estimation f_0 which can reduce the necessary number of iterations; and the threshold range $[\lambda_{\max}, \lambda_{\min}]$.

The generalization to full parallax is straightforward by successively implementing the basic algorithm along the horizontal and vertical camera axes. A more computationally-efficient alternative, referred to as hierarchical reconstruction [13], implements the reconstructions in a specific order, aimed at reducing the maximum disparity between input views after each iteration, thus reducing the required number of shearlet transform scales and the related processing time.

III. Accelerated Processing

The method in Section II is applicable for any EPI. A preferable solution would use the same set of parameters for every EPI and run the reconstructions independently. One can select optimal thresholding parameters for the reconstruction algorithm as $[\lambda_{\max}^{\text{opt}}, \lambda_{\min}^{\text{opt}}]$ and fix a common number of iterations N for all EPIs. In this case, the computation time linearly depends on the number of EPIs and the fixed number of iterations. By distributing the required computations equally between multiple GPUs, one can achieve the fastest computational time for this independent processing. Further acceleration can be achieved by speeding up the iterative algorithm itself and by utilizing similarities between EPIs.

A. Faster Convergence by Double Overrelaxation

In this section we propose a modification of the main iterative algorithm aimed at its faster convergence. As presented in (2), the convergence is controlled by the parameter α_n, which was originally designed to provide stability for varying content for the price of increased computations [13]. As a computationally less expensive alternative, here we propose another update mechanism, based on the so-called double overrelaxation (DORE), similar to the one presented in [18]. Assume the light field EPI matrices are reordered in column vector form and assume the parameter $\alpha_n = \alpha$ is fixed. The thresholding operation

$$\hat{f}_n = S^*(T_{\lambda_n}(S(f_n + \alpha(g - M \odot f_n)))) \quad (3)$$

is followed by a two-step overrelaxation

$$\tilde{f}_n = \hat{f}_n + \beta_1(\hat{f}_n - f_{n-1})$$
$$\beta_1 = \frac{(g - \hat{f}_n)^\top H(\hat{f}_n - f_{n-1})}{(\hat{f}_n - f_{n-1})^\top H(\hat{f}_n - f_{n-1})}, \quad (4)$$
$$f_{n+1} = \tilde{f}_n + \beta_2(\tilde{f}_n - f_{n-2})$$
$$\beta_2 = \frac{(g - \tilde{f}_n)^\top H(\tilde{f}_n - f_{n-2})}{(\tilde{f}_n - f_{n-2})^\top H(\tilde{f}_n - f_{n-2})}, \quad (5)$$

where $H \in R^{N^2 \times N^2}$ is a diagonal matrix containing the elements of the measuring matrix M along its main diagonal. For additional stability we clamp the values of $\beta_1, \beta_2 \in [0, 1]$. The role of the double over-relaxation is to tackle potential instabilities by keeping the next iteration anchored to the previous iterations. Equations (4) and (5) provide closed-form solutions for the respective line search problems $\beta_1 = \underset{\beta}{\text{argmin}} ||H(g - (\hat{f}_n + \beta(\hat{f}_n - f_{n-1}))||^2$ and $\beta_2 = \underset{\beta}{\text{argmin}} ||H(g - (\tilde{f}_n + \beta(\tilde{f}_n - f_{n-2}))||^2$. This leads to finding an optimal linear combination between consecutive solutions such that the error is minimized over the given samples defined by the matrix H.

B. Color Spaces and Guided Colorization

A trivial approach is to convert the RGB color channels into YUV colour space and process EPIs there, while expecting significantly less energy in the U and V colour channels. Specifically, we apply reversible color transform (RCT [19]) without any quantization of values, i.e.

$$\begin{cases} Y = (R + 2 * G + B)/4 \\ U = B - G \\ V = R - G \end{cases}.$$

Usually, the spatial information in U and V channels is highly redundant for natural images. Therefore, in the case of processing in YUV colour space with given N number of iterations, we reconstruct Y channel with N iterations and U, V channels with $N/2$ iterations. Compared with the reconstruction in RGB colour space, the overall number of iterations is reduced from $3N$ to $2N$.

Furthermore, we investigate the possibility of applying the fully reconstructed Y-channel EPI as a guide in reconstructing R, G and B color channels from their decimated EPI versions. This type of problem can be solved by methods previously developed for image colorization [20], [21]. It has been also shown that colorization can be considered as a particular case of the more general problem of alpha matting [22]. Specifically, we adopt the so called closed-form alpha matting algorithm proposed in [23] and modify it for the purpose of reconstructing color EPIs.

Following the notations in Section II, we denote the targeted EPI color channel and its decimated version by f and g respectively. Let us denote also the reconstructed Y-channel EPI by E. Then, the targeted color channel pixels f_i are modelled as a linear function of the known (i.e. guiding) image pixels E_i,

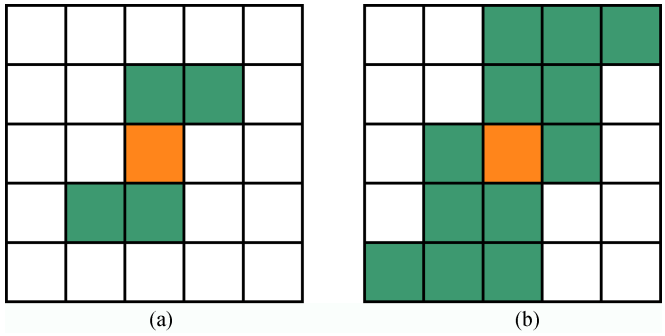


Fig. 2. (a) Proposed window (green) for modelling guidance with respect to reference pixel (orange). (b) Neighbourhood (green) for forming Laplacian matrix entry with respect to reference pixel (orange).

within a small window w

$$f_i \approx aE_i + b, \forall i \in w.$$

For natural images, typically, the small window has been assumed to be a square window of 3×3 pixels around the reference pixel. For the case of EPI, we propose to use a different window to leverage the directional information presented in EPI. We show the proposed shape in Fig. 2(a).

The cost function minimization problem can be formulated as follows

$$J(f,a,b) = \sum_{j=1}^{N^2} \left(\sum_{i \in w_j} (f_i - a_j E_i - b_j)^2 + \varepsilon a_j^2 \right),$$

where the regularisation term εa_j^2 is added for numerical stability. In [23], it has bee shown that an equivalent minimization problem can be formulated using *matting Laplacian* matrix Λ, which removes the need to identify a and b

$$J(f) = \min_{a,b} J(f,a,b) \sim J(f) = f^T \Lambda f,$$

where the entries of the matrix $\Lambda \in R^{N^2 \times N^2}$ are calculated as follows

$$\Lambda(i,j) =$$

$$\sum_{k|(i,j)\in w_k} \left(\delta_{ij} - \frac{1}{\aleph_k} \left(1 + \frac{1}{\frac{\varepsilon}{\aleph_k} + \sigma_k^2} (E_i - \mu_k)(E_j - \mu_k) \right) \right).$$

In the above equation, δ_{ij} denotes the Kronecker delta, and $\aleph_k$, μ_k, σ_k^2 denote the cardinality, the mean and the variance of the window w_k respectively. For the entry $L(i,j)$, the summation is done over all windows w_k which contain pixels with indices i and j. For a reference pixel at position i and for our choice of window shape, the pixel positions j are shown in Fig. 2(b). Given the true colors at the decimated EPI g, the problem is reformulated as

$$\text{minimize } f^T \Lambda f, \text{ s.t. } Hf = g,$$

where H is the diagonally-arranged measurement matrix M. The so-formulated problem is solved using the conjugated gradient method.

C. Group Processing of Similar EPIs

Previously, we have presented an attempt to accelerate the basic algorithm utilizing similarities between EPIs [24]. The method suggested constructing a tree, which defines the order of processing depending on similarity between EPIs. In the constructed tree, each node corresponds to an EPI. The tree is constructed by comparing EPIs for their similarity in terms of l^2 norm and consecutively connecting the most similar pairs of EPIs. Iterating over all EPIs, one obtains a connected graph. Then, the processing is performed from top to bottom, and the EPI being processed uses the reconstructed EPI at its parent node as an initial estimate. Our hypothesis was that the reconstruction over the graph would allow for adaptively choosing the number of iterations for each EPI depending on the similarity to its initial estimate. This approach heavily depends on the threshold defining similarity between EPIs and setting the same reconstruction parameters for different datasets is problematic. Therefore, in this article we adopt a more systematic approach toward exploring the EPI similarities, which would allow an easier tuning of algorithm parameters. First, we consider grouping of similar EPIs, done by comparing l^2 distances between EPIs against a predefined threshold t_s. Having the EPIs organized in groups, we fully reconstruct the average EPI over each group and use it as a guidance map to reconstruct the other EPIs in the group by the approach proposed in Section III-B.

D. Initialization by Wavelet Transform

The redundancy between EPIs can be regarded as redundancy in the vertical direction in the given multi-perspective images. Instead of local grouping of similar EPIs as in Section III-C, we consider the alternative of decorrelating the vertical image lines by a fixed transform, e.g., a wavelet transform. Namely, a wavelet transform is performed on $E_i^c(\cdot,\cdot)$ along the i axis which is equivalent to performing 1D wavelet transform vertically on every input image $I_k^c(i,\cdot)$ along i axis. By performing L level of 1D wavelet transform between EPIs $E_i^c, i = 1,...,p$, we expect to split them into EPIs with small-magnitude detail coefficients and $\tilde{E}_i^c, i = 1,...,p/2^L$ EPIs with higher-magnitude approximation coefficients. The approximation coefficients gather most of the information of the original set of EPIs. The reconstruction is then applied directly on EPIs formed by wavelet transform approximation coefficients. The obtained set of densely sampled EPIs $\tilde{E}_i^d, i = 1,...,p/2^L$ contain a good amount of directional structures (more global ones), however, they still require further processing to obtain desirable quality of reconstruction (add details from the original set of EPIs). Therefore, the inverse wavelet transform can be applied on the reconstructed EPIs of the approximation coefficients with an appropriate padding with zeros corresponding to detail coefficients. The obtained set of EPIs $\tilde{I}_i^d, i = 1,...,p$ is used as an initial estimate for reconstruction of the original input E_i^c EPIs by performing additional processing by the modified basic algorithm. The processing times for the two steps can be set independently. The flowchart of the approach is shown in Fig. 3.

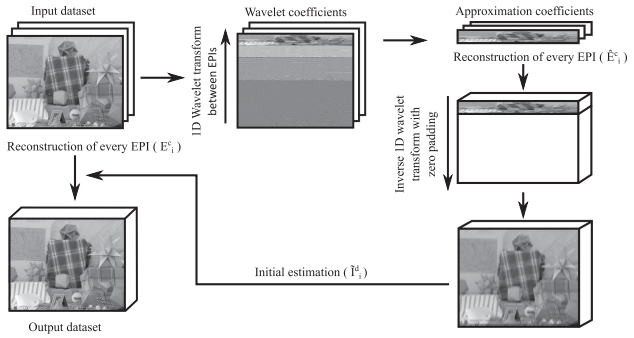


Fig. 3. Reconstruction flowchart using wavelet transform approximation coefficient as an initial estimation.

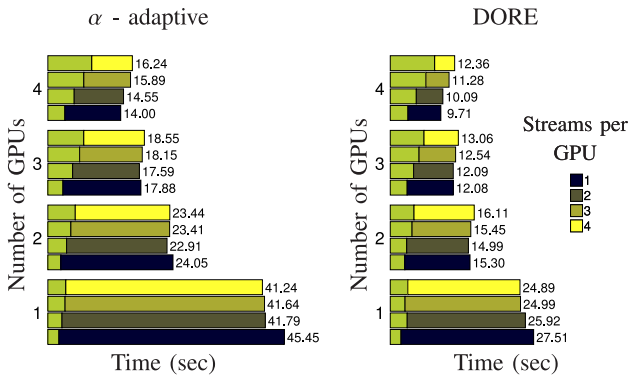


Fig. 4. Computational time required to perform 50 iterations of α-adaptive algorithm (left) and $DORE$ (right) on 100 EPIs using different parallelization between GPUs. Light green color represents the necessary time for initialization of the algorithm. For resolution 256×512 using more than one stream per GPU doesn't provide acceleration.

IV. Experimental Evaluation

A. Algorithm Implementation

We have implemented the core reconstruction algorithms, on a GPU using CUDA Toolkit [25]. Since the shearlet transform is a translation invariant transform, it can be efficiently computed using the Fast Fourier Transform (FFT). In our case, we used the cuFFT library to get an FFT implementation on the GPU [26]. For generating our experimental results we have used a system consisting of four Nvidia GeForce GTX Titan X GPUs. The computational time for reconstructing an EPI mainly depends on the number of overall iterations that have to be performed. In order to achieve fastest computation for a given set of input EPIs with a corresponding number of iterations, the set has been distributed between GPUs such that the overall number of iterations that has to be performed are approximately equal for each GPUs. On the level of one GPU, the whole iterative processing is performed independently from other GPUs. Depending on the size of the processed EPI, we get different occupation of GPU kernels at a time. We consider EPIs with the size of 256×512 processed with the shearlet transform at 5 scales using algorithm presented in Sections II, III-A. In both cases, 100 EPIs are processed. The computational times are presented in Fig. 4. Note that reconstructing in the case of 256×512 with S^5 transform, only one process per GPU is sufficient for both algorithms, while DORE is significantly faster.

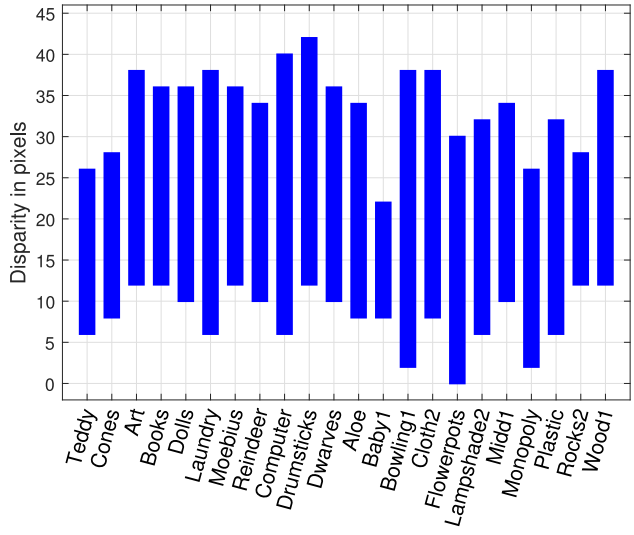


Fig. 5. Illustration of the disparity ranges $[d_{\min}, d_{\max}]$ between adjacent views for input dataset used in this paper (obtained from ground true disparity maps).

B. Evaluation

In our comparative tests, we have used datasets presented in [27], [28] for horizontal parallax and in [29] for full parallax datasets. In overall, 22 horizontal-parallax datasets and two full-parallax datasets of various depth and spatial content have been used. In all experiments, the input data is formed by every second view of the test dataset. The other views form the reference. The algorithm performance has been evaluated by comparing the difference between the reconstructed and the reference views in terms of PSNR (dB). $J = 5$ scale levels have been considered for the shearlet transform (S^5) which corresponds to an intermediate view reconstruction, where the maximum disparity between adjacent views is in the range of $[0, 32]$ pixels. The exact disparity ranges of each scene as obtained from the ground truth disparity maps are shown in Fig. 5. In order to shift the available (existing) disparity ranges to $[0, d_{\max} - d_{\min}]$, for each input dataset we perform first a horizontal shearing by $-d_{\min}$ on all EPIs. After reconstructing $d_{\max} - d_{\min} - 1$ intermediate views, a shearing by $d_{\min}/(d_{\max} - d_{\min})$ is applied to return the imagery to the original disparity range. In general, we evaluate algorithms presented in Section III for different number of iterations in order to compare trend of convergence speed of different algorithms in average for all datasets.

In our first comparative test, we present the average reconstruction quality for the algorithm modification based on DORE with $\alpha = 20$ and the original α adaptive algorithm. The comparison is done for $5, 10, 15, 20, 30, 50, 75, 100$ iterations per EPI. The results for the horizontal parallax datasets are shown in Fig. 6, while the reconstruction results for the full-parallax datasets are shown in Fig. 7. The DORE algorithm provides faster convergence for all datasets and results in better quality enjoying also faster processing. In all subsequent experiments, we use the DORE algorithm with $\alpha = 20$, referring to it as the reference algorithm.

Next, we aim at quantifying the performance of the colorization algorithm. For the horizontal-parallax datasets, we present

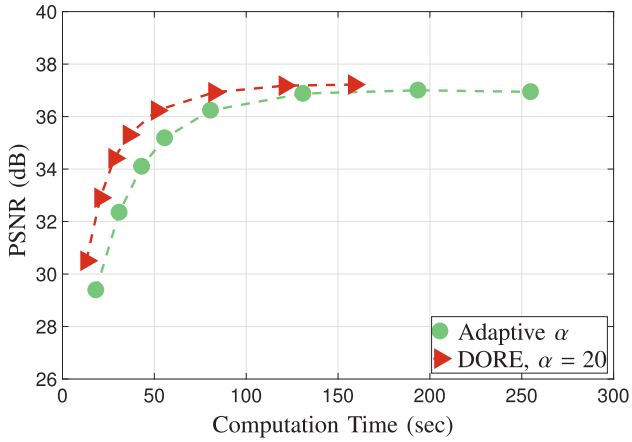


Fig. 6. Comparison of average performance between basic algorithm with adaptive selection of the parameter α and DORE with fixed $\alpha = 20$ for the horizontal-parallax datasets.

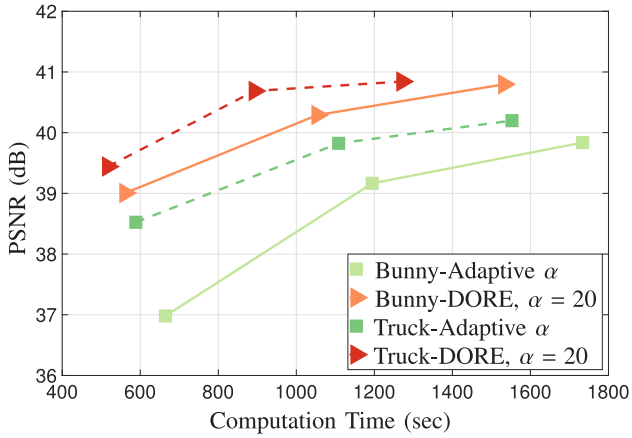


Fig. 7. Comparison of performance between basic algorithm with adaptive selection of the parameter α and DORE with fixed $\alpha = 20$ for the full-parallax datasets.

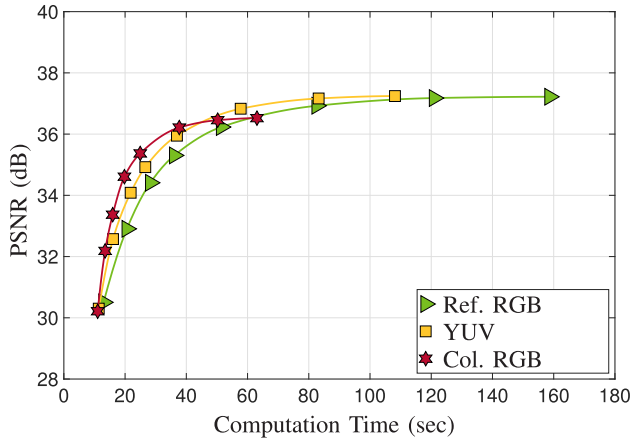


Fig. 8. Comparison between reference algorithm (RGB), YUV, and colorization of RGB for the horizontal-parallax datasets.

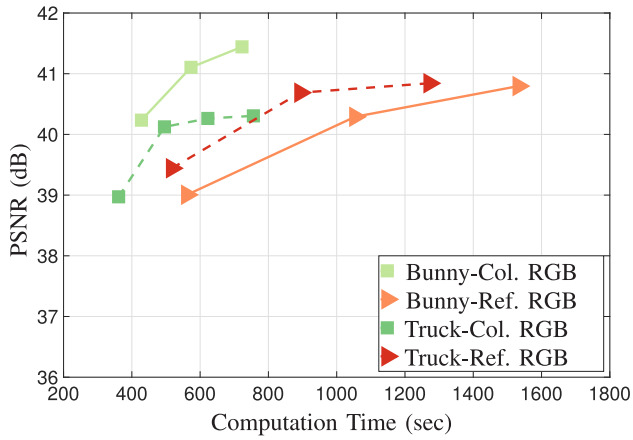


Fig. 9. Comparison between reference algorithm (RGB), and colorization of RGB for the full-parallax datasets.

the trend in reconstruction quality for different number of iterations, see Fig. 8. The reference algorithm reconstructs the three color channels, R, G, and B in an equal number of iterations, while in YUV, priority is given to the Y channel, which is processed twice longer than the U and V channels. In the case of colorization, an average intensity channel is formed as $Y = (1/3) * (R + G + B)$ and fully reconstructed in varying number of iterations, then each of the color channels is reconstructed by colorization using the reconstructed Y channel as a guidance. As can be seen in the figure, the prioritized processing brings some improvement over the reference algorithm and the algorithm based on colorization is significantly faster as it processes a single channel only. All three algorithms saturate in performance, which means that after some number of iterations, no quality improvement is achieved. The colorization algorithm saturates at lower level, which indicates that the structural differences in the three color channels have not been fully reconstructed in the averaged intensity channel. The reached values after 100 iterations for each of the three algorithms are the following: $RGB - 37.22$ dB, $YUV - 37.24$ dB, $Col.RGB - 36.22$ dB. These results suggest that the colorization algorithm is preferable in case of limited computation time, since it converges faster, while for the case of better computing resources, the best quality is achieved by the YUV color space processing.

Fig. 9 presents the results for the full-parallax datasets. For the $Bunny$ dataset, the colorization shows a significant improvement both in terms of time and quality, while for the $Truck$ dataset, the results are in agreement with the average result over the horizontal-parallax datasets.

Fig. 8 presents the average results over the whole group of test scenes. The result in terms of rate of convergence vary substantially for the individual test scenes. In order to further analyse the algorithm based on colorization, we look at the saturation points for the reference algorithm and the colorization algorithm for each individual dataset. The two algorithms are run for increasing number of iterations and saturation points are estimated at the iteration where further improvement is negligible. Denote by $E(k), T(k), k = 1, \ldots$ the quality level (e.g., PSNR), and the corresponding time in seconds for the running iteration k. We define $k_{\max} = \arg\max_k E(k)$ and define the saturation point as $k_{sat} = \arg\max_k \left(\frac{E(k+1)-E(k)}{E(k_{\max})} < 0.0015 \right)$. The idea is illustrated in Fig. 10 for the Teddy dataset, where the saturation points are given by the circles around the corresponding iterations.

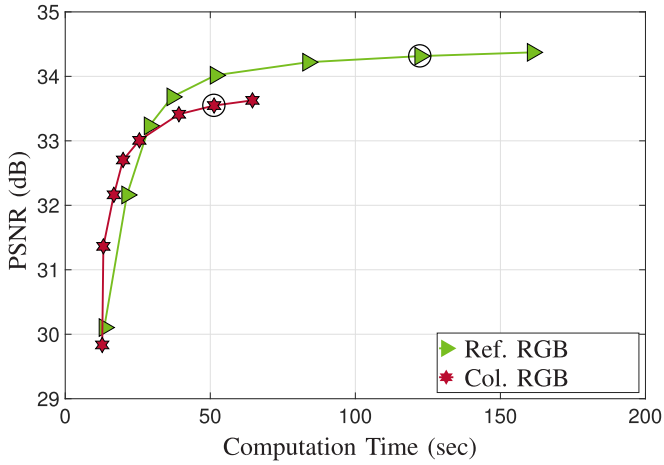


Fig. 10. Saturation points for the reference and colorization algorithms for the Teddy dataset. The obtained saturation points are illustrated by black circles.

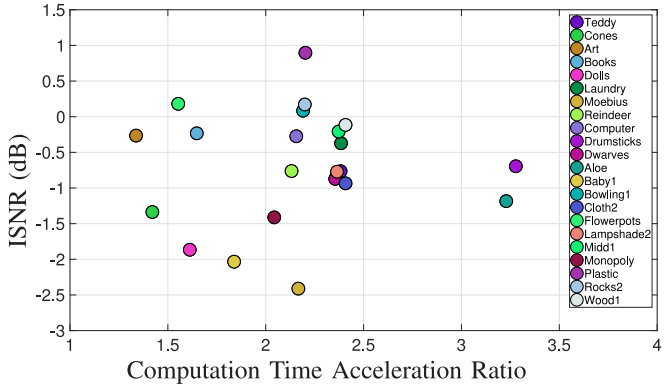


Fig. 11. Comparison of saturation points of the colorization and reference algorithms for the horizontal-parallax test datasets.

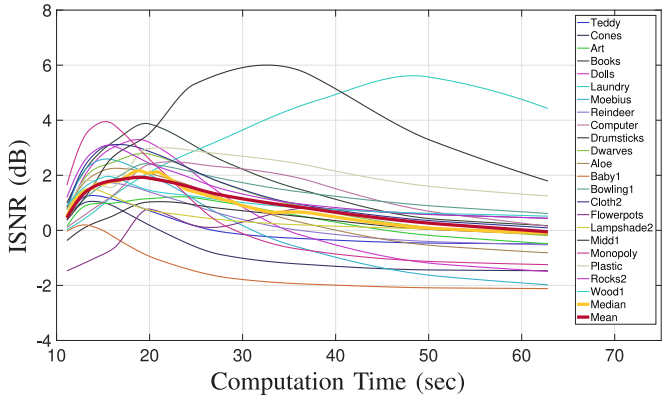


Fig. 12. ISNR of colorization versus reference algorithm for individual sequences.

Having found two saturation points per dataset, one for the reference and one for the colorization algorithm, one can compare them in terms of quality variation (ISNR) and time acceleration ratio. In other words, we compare the best achievable quality per dataset for the two algorithms versus the time acceleration ratio it brings. Fig. 11 presents this comparison over the

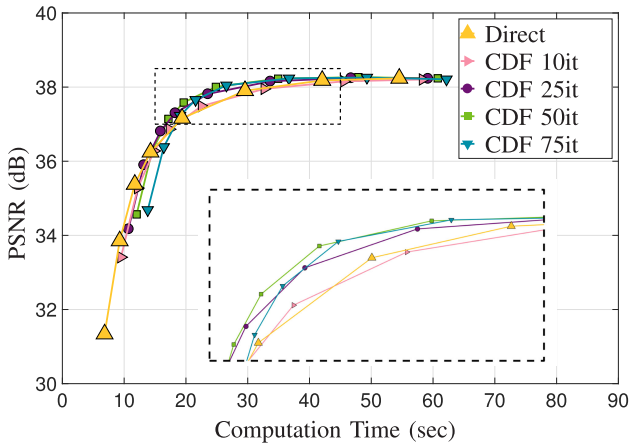


Fig. 13. Comparison of the reconstruction trends depending on number of iterations used for obtaining initial estimation in method utilizing wavelet transform. In the legend of the figure presented corresponding number of iterations used for processing initial estimations.

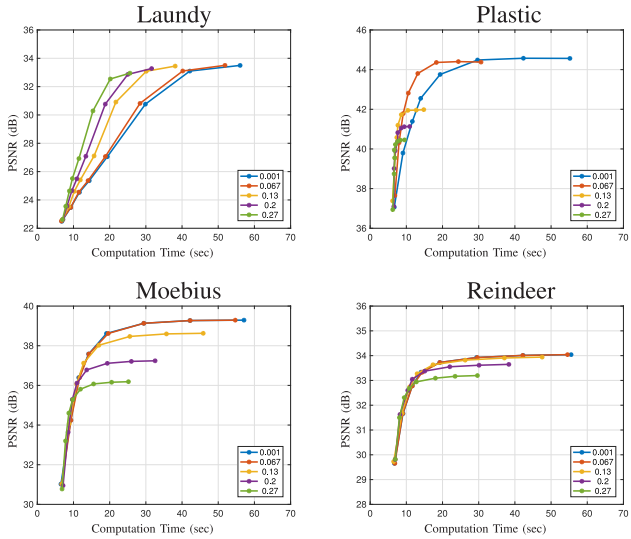


Fig. 14. Comparison of the reconstruction trends depending on different thresholding value in method utilizing groups formed based on similarity of the EPIs. In the legend of the figure presented corresponding thresholding values.

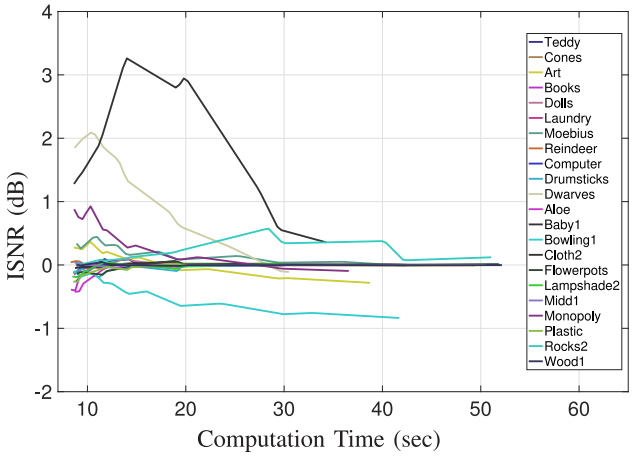


Fig. 15. Average reconstruction performance for the method utilizing grouping based on similarity between EPIs for different datasets for fixed thresholding value.

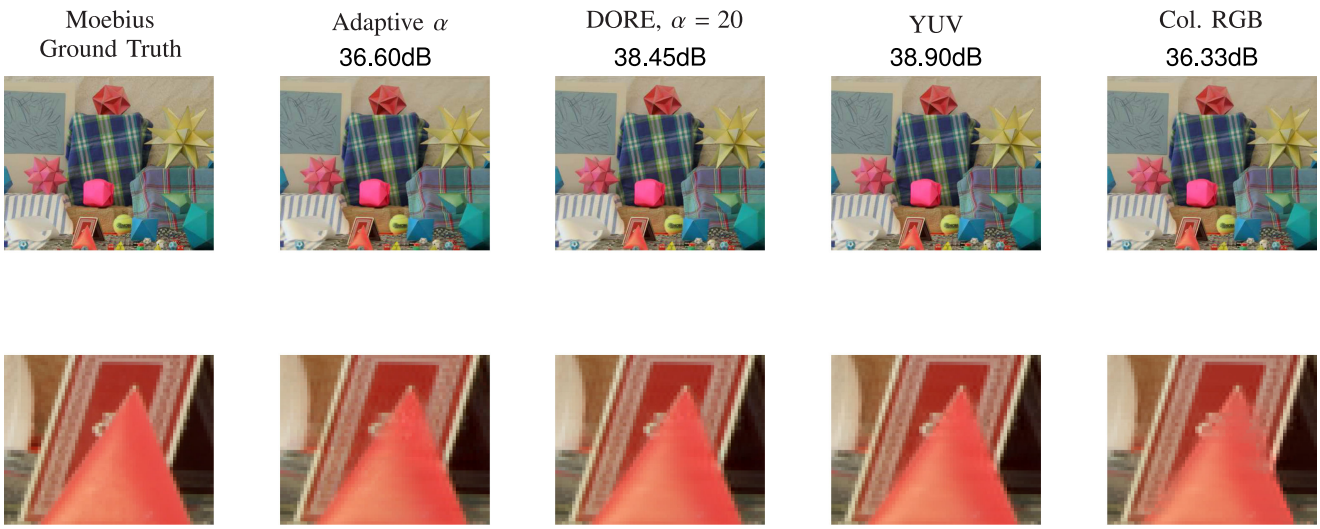


Fig. 16. Example of synthesized views for datasets $Moebius$ using different algorithms for approximately same computation time. Presented algorithms are used for accelerated color space reconstruction.

horizontal-parallax datasets. In the figure, each dot represents one dataset, the $x-$ axis represents the relative time acceleration achieved by the colorization algorithm versus the reference one, and the $y-$ axis represents the improvement in the signal-to-noise ratio ISNR (dB). As seen in the figure, there are a few sets, where the acceleration in time comes together with improved quality, while for the majority of datasets, the acceleration is achieved for the price of reduced quality.

Another way to illustrate the performance of the colorization algorithm is to show the ISNR in comparison to the reference algorithm for the same time, using interpolation between iteration points. Fig. 12 gathers the performance for each individual dataset, along with the mean and median values. For short processing times, colorization is to be preferred as most of the sequences show positive ISNR values. As the processing time gets longer, the values cluster around the zero ISNR line (as shown also by the mean and median curves), while there are a few sequences still enjoying better performance with the colorization method and other sequences showing worsening results. Apparently, among the former group these are datasets with relatively simpler color and depth distributions.

To simplify the next experiments, we limit the comparison of algorithms exploring the inter-EPI similarities and decorrelation to comparing the results for the Y channel only, assuming that the RGB color channels can be efficiently reconstructed by the Y channel.

For the wavelet transform based acceleration approach, presented in Section III-D, we perform $L = 3$ levels of the CDF $9/7$ transform. Here, the issue is to find the best proportion between the processing time allocated for obtaining the initial estimate and the processing time allocated for refining the EPI reconstruction based on this initial estimate. In order to illustrate the trend in convergence, we perform experiments where the initial estimate is obtained by reconstructing the coarse wavelet coefficients with $10, 25, 50, 75$ iterations. The obtained initial estimate is then refined for the same number of iterations, as the direct (reference) algorithm. Fig. 13 depicts the trends. Naturally, the time needed for obtaining the initial estimates, shift the initial curve points to the right, e.g., the curve corresponding to 75 iterations allocated for getting the initial estimate is the rightmost in the figure. Then, the curves corresponding to wavelet-based initialization get better and saturate faster, with the case of 50 iterations for the initial estimate showing the best performance.

The algorithm based on grouping similar EPIs and processing them together as presented in Section III-C does not show consistent results for all datasets. This is to be attributed to the strong dependence on the threshold value, which determines which EPIs are sufficiently similar. In general, increasing the threshold value leads to increasing the size of formed groups and therefore decreases the computation time. However, the effect on quality is very much content-dependent. Results for several datasets with different threshold values are presented in Fig. 14. For the dataset $Laundry$, one can get significant acceleration, while e.g. for the dataset $Moebius$, the reconstruction quality is always inferior compared with the reference algorithm. Selecting one of the well-performing thresholds, i.e. the value of 0.067, one can get the performance for each individual dataset, as shown in Fig. 15.

Some examples of synthesized views are presented in Figs. 16 and 17 with the corresponding quality in terms of PSNR. The DORE-based algorithm provides consistently better convergence compared to the original method [13]. The colorization-based algorithm achieves good reconstruction results when the Y channel manages to get the important structure of the scene (edge information) existing in R, G and B channels. For the particular scene $Laundry$, the algorithm based on wavelet transform provides better convergence compared to the reference RGB algorithm.

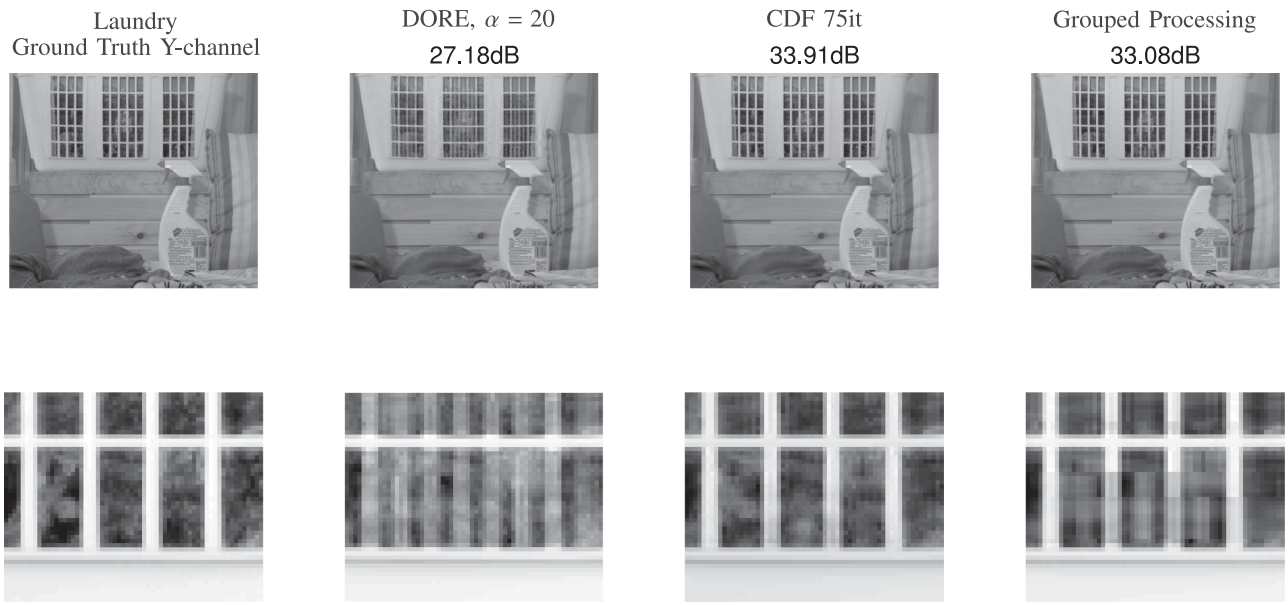


Fig. 17. Example of synthesized views for only Y-channel of dataset $Laundry$ using different decorrelation algorithms for approximately same computation time.

V. Conclusion

In this article, we have addressed the problem of accelerating the DSLF reconstruction algorithm, which originally uses sparse camera views and works on each EPI independently by employing regularized iterative reconstruction in shearlet transform domain. In order to speed up the algorithm, we proposed modifications in three categories. First, we aimed at improving the algorithm itself by using double relaxation in the iterative procedure. Second, we explored the similarities between color channels within the same EPI in the flavor of colorization based approaches. Third, we aimed at avoiding redundant processing and reducing the overall reconstruction time through exploiting similarities between EPIs. Furthermore, our implementation employed GPUs allowing for an efficient parallelized computation of the iterative procedure and the underlying shearlet transform.

We have generated experimental results on a wide set of test sequences and analyzed the performance of the considered approaches. The new reconstruction method based on double overrelaxation shows better convergence speed in comparison with the original algorithm. We favor the use of colorization as the approach catches well the color dependences in natural images. The benefit of using similarities between EPIs is very much content dependent. The wavelet based approach shows a marginal improvement in terms of convergence rate, which is still worth employing. As of the algorithm based on grouping of similar EPIs and group processing, it provides acceleration only for scenes where significant amount of EPIs are similar.

The modifications employ structured similarities within EPI and between EPIs were integrated within the DSLF reconstruction algorithm. However, they are perfectly applicable also in other LF image processing algorithms, where DSLF reconstruction is not the main goal. Such potential applications include LF depth estimation, compression, segmentation, and matting.

References

[1] N. S. Holliman, N. A. Dodgson, G. E. Favalora, and L. Pockett, "Three-dimensional displays: a review and applications analysis," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 362–371, Jun. 2011.

[2] B. Javidi and F. Okano, *Three-Dimensional Television, Video, and Display Technologies*. New York, NY, USA: Springer, 2002.

[3] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.

[4] Z. Lin and H.-Y. Shum, "A geometric analysis of light field rendering," *Int. J. Comput. Vision*, vol. 58, no. 2, pp. 121–138, 2004.

[5] C. Fehn, "Depth-image-based rendering (DIBR) compression and transmission for a new approach on 3D-TV," *Proc. SPIE Conf. Stereoscopic Displays and Virtual Reality Systems XI*, vol. 5291, pp. 93–104, 2004.

[6] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen, "Unstructured lumigraph rendering," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (SIGGRAPH '01 Series). New York, NY, USA: ACM, 2001, pp. 425–432. [Online]. Available: http://doi.acm.org/10.1145/383259.383309

[7] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR '06 Series) vol. 1. Washington, DC, USA: IEEE Computer Society, 2006, pp. 519–528. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2006.19

[8] R. Szeliski *et al.*, "A comparative study of energy minimization methods for Markov random fields with smoothness-based priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1068–1080, Jun. 2008.

[9] T. Pock, D. Cremers, H. Bischof, and A. Chambolle, "Global solutions of variational models with convex regularization," *SIAM J. Imag. Sci.*, vol. 3, no. 4, pp. 1122–1145, 2010.

[10] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 73:1–73:12, Jul. 2013.

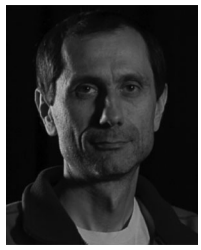
[11] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous fourier domain," *ACM Trans. Graphics*, vol. 34, no. 1, 2014, Art. no. 12.

[12] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Image based rendering technique via sparse representation in Shearlet domain," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 1379–1383.

[13] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light field reconstruction using Shearlet transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.
[14] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, *Depth Estimation Reference Software (DERS) 5.0*, ISO/IEC JTC1/SC29/WG11 M, vol. 16923, 2009.
[15] M. Tanimoto, T. Fujii, and K. Suzuki, *Reference Software of Depth Estimation and View Synthesis for FTV/3DV*, ISO/IEC JTC1/SC29/WG11 M, vol. 15836, 2008.
[16] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.
[17] C.-K. Liang, Y.-C. Shih, and H. Chen, "Light field analysis for modeling image formation," *IEEE Trans. Image Process.*, vol. 20, no. 2, pp. 446–460, Feb. 2011.
[18] K. Qiu and A. Dogandzic, "Double overrelaxation thresholding methods for sparse signal reconstruction," in *Proc. 44th Annu. Conf. Inf. Sci. Syst.*, Mar. 2010, pp. 1–6.
[19] Y. Chen and P. Hao, "Integer reversible transformation to make JPEG lossless," in *Proc. 7th Int. Conf. Signal Process.*, vol. 1, 2004, pp. 835–838.
[20] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 689–694, 2004.
[21] Q. Luan, F. Wen, D. Cohen-Or, L. Liang, Y.-Q. Xu, and H.-Y. Shum, "Natural image colorization," in *Proceedings of the 18th Eurographics Conference on Rendering Techniques* (ser. EGSR'07 Series), Aire-la-Ville, Switzerland: Eurographics Association, 2007, pp. 309–320. [Online]. Available: http://dx.doi.org/10.2312/EGWR/EGSR07/309-320
[22] J. Wang and M. F. Cohen, "Image and video matting: A survey," *Found. Trends. Comput. Graph. Vis.*, vol. 3, no. 2, pp. 97–175, Jan. 2007. [Online]. Available: http://dx.doi.org/10.1561/0600000019
[23] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 228–242, Feb. 2008.
[24] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Tree-structured algorithm for efficient shearlet-domain light field reconstruction," in *Proc. IEEE Global Conf. Signal Inf. Process.*, IEEE, 2015, pp. 478–482.
[25] J. Nickolls, I. Buck, M. Garland, and K. Skadron, "Scalable parallel programming with cuda," *Queue*, vol. 6, no. 2, pp. 40–53, 2008.
[26] V. Podlozhnyuk, "FFT-based 2D convolution," NVIDIA white paper, NVIDIA Corporation, Santa Clara, CA, USA, 2007, p. 32.
[27] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, vol. 1, Jun. 2003, pp. I–195–I–202.
[28] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2007, pp. 1–8.
[29] V. Vaish and A. Adams, The (New) Stanford Light Field Archive. 2008. [Online]. Available: http://lightfield.stanford.edu



Suren Vagharshakyan (M'15) received the M.Sc. degree in mathematics from Yerevan State University, Yerevan, Armenia, in 2008. He has been working toward the Ph.D. degree in the Department of Signal Processing, Tampere University of Technology, Tampere, Finland, since 2013. His research interests include the area of light field capture and reconstruction.



Robert Bregovic (M'97) received the Dipl.Ing. and M.Sc. degrees in electrical engineering from the University of Zagreb, Zagreb, Croatia, in 1994 and 1998, respectively, and the Dr.Sc. (Tech.) degree (with honors) in information technology from Tampere University of Technology, Tampere, Finland, in 2003. From 1994 to 1998, he was with the Department of Electronic Systems and Information Processing, Faculty of Electrical Engineering and Computing, University of Zagreb. Since 1998, he is with the Laboratory of Signal Processing, Tampere University of Technology. His research interests include the design and implementation of digital filters and filterbanks, multirate signal processing, and topics related to acquisition, processing, modeling, and visualization of 3-D content.



Atanas Gotchev (M'95) received the M.Sc. degree in radio and television engineering in 1990 and in applied mathematics in 1992, and the Ph.D. degree in telecommunications in 1996 all from the Technical University of Sofia, Sofia, Bulgaria, and the D.Sc. (Tech.) degree in information technologies from the Tampere University of Technology, Tampere, Finland, in (2003). He currently is a Professor at the Laboratory of Signal Processing and Director, Centre for Immersive Visual Technologies, Tampere University of Technology. His research interests include sampling and interpolation theory, and spline and spectral methods with applications to multidimensional signal analysis. His recent work concentrates on algorithms for multisensor 3-D scene capture, transform-domain light-field reconstruction, and Fourier analysis of 3-D displays.

Comparison and Evaluation of Light Field Image Coding Approaches

Irene Viola, Martin Řeřábek, and Touradj Ebrahimi, *Member, IEEE*

***Abstract*—The recent advances in light field imaging, supported among others by the introduction of commercially available cameras, e.g., Lytro or Raytrix, are changing the ways in which visual content is captured and processed. Efficient storage and delivery systems for light field images must rely on compression algorithms. Several methods to compress light field images have been proposed recently. However, in-depth evaluations of compression algorithms have rarely been reported. This paper aims at the evaluation of perceived visual quality of light field images and at comparing the performance of a few state-of-the-art algorithms for light field image compression. First, a processing chain for light field image compression and decompression is defined for two typical use cases, professional and consumer. Then, five light field compression algorithms are compared by means of a set of objective and subjective quality assessments. An interactive methodology recently introduced by authors, as well as a passive methodology is used to perform these evaluations. The results provide a useful benchmark for future development of compression solutions for light field images.**

***Index Terms*—Image coding, image compression, light field, objective quality evaluation, subjective quality evaluation.**

I. Introduction

THE idea that light, just as electromagnetism, can be interpreted as a field was first proposed by Michael Faraday in 1846. The concept was subsequently formalized by Andreī Gershun, who coined the term *light field* (LF) in his book on radiometric properties of light in 3D space [1].

One way to represent the LF is to describe the radiance along the light rays in a 3D space with constant illumination. This can be achieved by using the *plenoptic function*, which was first introduced in 1991 by Adelson and Bergen [2]. More particularly, the plenoptic function $\mathcal{L}$ describes the intensity of the light rays passing through every possible point in space (V_x, V_y, V_z) at every possible angle (θ, ϕ), wavelength λ, and time t, represented as follows:

$$\mathcal{L} = \mathcal{L}(\theta, \phi, \lambda, t, V_x, V_y, V_z). \tag{1}$$

Assuming a 3D region free of occlusions at a single time instance and considering the fact that radiance along rays remains constant in a free space, the above 7D plenoptic function can be further simplified into a *4D light field* function [3]. Such 4D function, representing a set of light rays, can be parametrized as an intersection of rays with two planes: uv describing the rays position in aperture (object) plane and xy describing the rays position in image plane.

$$\mathcal{L} = \mathcal{L}(u, v, x, y). \tag{2}$$

The parametrized LF function, further referred to as 4D LF, can be considered as a collection of perspective images of the xy plane, each observed from a position on the uv plane.

In the past, LF images have been mainly used to navigate through 3D scenes. Recently, their applications have expanded thanks to the creation of commercially available devices, such as Raytrix or Lytro cameras, as well as development of new visualization devices capable of properly coping with LF images. However, due to the enhanced features that LF imaging offers, a vast amount of data is created during the acquisition step. Therefore, it is necessary to find an efficient way to compress LF images for transmission and storage.

Currently available techniques to capture and visualize LF images determine two general approaches for LF image compression. A general diagram of workflow for LF image acquisition and visualization is depicted in Fig. 1. The first coding approach assumes that the raw sensor data obtained during the acquisition step is compressed directly with minimal signal preprocessing such as demosaicing or devignetting (point A in Fig. 1). The actual format of raw data strongly depends on the exact acquisition device, e.g., a lenslet based hand-held camera, a multi-camera array, or a multi-view plus depth acquisition device. Often, extensive post-processing of the decompressed LF image is necessary prior to its visualization. Furthermore, additional metadata about the captured scene and acquisition device, e.g., camera and color calibration data, is needed to properly process and visualize the LF image. The second coding approach considers creation of a 4D LF representation of LF image prior to compression (point B in Fig. 1). As mentioned above, the 4D LF represents a collection of perspective images, which can be visualized without a need for acquisition related metadata or post-processing. Process of creation of 4D

Manuscript received December 1, 2016; revised April 11, 2017 and June 19, 2017; accepted August 3, 2017. Date of publication August 15, 2017; date of current version October 23, 2017. This work was conducted in the framework of the Swiss National Foundation for Scientific Research (FN 200021_159575) project Light field Image and Video coding and Evaluation (LIVE) and also in the framework of ImmersiaTV under the European Unions Horizon 2020 research and innovation programme (Grant 688619) funded by Swiss State Secretariat for Education, Research and Innovation SERI. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Yebin Liu. *(Corresponding author: Irene Viola.)*

The authors are with the Multimedia Signal Processing Group, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne 1015, Switzerland (e-mail: irene.viola@epfl.ch; touradj.ebrahimi@epfl.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2740167

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

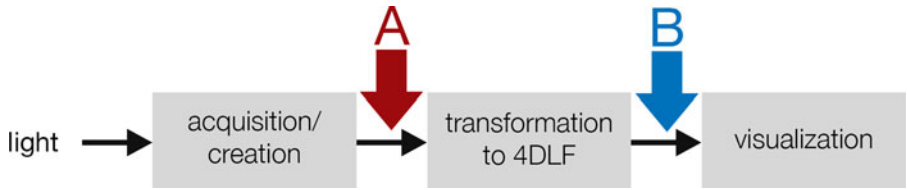


Fig. 1. General acquisition and display pipeline for LF images.

LF from the raw sensor data also strongly depends on the exact acquisition device.

In the context of general LF image manipulation, one can think of two specific use cases related to either of the two coding approaches defined above. On the one hand, professional photographers, operators, and artists may benefit from LF image acquisition technologies, since they allow for greater flexibility in terms of optimal parameters selection after capture. For example, an erroneous selection of focal plane in a scene may lead to several retakes and thus to greater costs. Other features, such as change of point of view or zoom, may dramatically impact the way scenes are captured. In this case, it is of paramount importance that key factors in the acquisition, such as white balance, color, and exposition, are not altered in the compression step, and that acquisition metadata is stored to be used during post-processing.

Consumers, on the other hand, may turn to LF imaging when looking for an enhanced experience to capture a special moment. Ability to change zoom, perspective, and focus in a simple and intuitive way without the need for expensive post-processing software, is in line with the interactivity already seen in applications like Instagram, in which users can modify the appearance of the captured scenes with predefined filters. In this case, the fidelity to the acquisition parameters is less important. However, the resulting image should not be too large and ready to be visualized and shared in devices with limited resources.

In this paper, we compare two main coding approaches to compress LF images through a set of objective and subjective quality evaluations. More specifically, five different compression algorithms suitable to either of predefined use cases are described, investigated, and evaluated. A standard PSNR metric, adjusted to the properties of LF images, is used as objective quality metric.

In the past, the authors have used single viewpoint image visualization to assess the visual quality of LF contents [4]. However, this type of assessment is suboptimal for two main reasons. The first concerns the number of stimuli needed to perform the tests, which increases significantly once multiple viewpoints and refocusing points are selected (in case of the test conducted in this paper, 720 stimuli were tested). The increased number of stimuli leads to strain and fatigue for the subjects and is generally not recommended. Secondly, analysis of variance performed on the results obtained from different viewpoints showed that no significant difference can be found between scores, meaning that subjects would rate different viewpoints from the same content in a similar way. Thus, evaluating different viewpoints leads to a lengthy, more costly and less efficient test, with no visible gain. For the reasons mentioned above, two methodologies have been selected to subjectively assess the visual quality of LF image contents. The first methodology was recently introduced by the authors and allows for interaction with the displayed content [5]. The second methodology passively shows the different viewpoints composing the LF content, in order to ensure that all users see and rate the exact same content. The two methodologies have been analysed by the authors and were found to be highly correlated [6]. The two methodologies are described in details and applied to the problem of assessing the visual quality of the five compression algorithms.

The remainder of the paper is organized as follows. Section II presents the state of the art in related areas of LF imaging, specifically LF image acquisition, compression, and quality evaluation. Section III discusses in details the two approaches examined in the paper. Section IV presents our experiments in details, whereas Section V exposes and comments the results. Section VI concludes the paper.

II. State of the Art

This section briefly reviews the state of the art in LF imaging. First, a quick overview of LF image acquisition techniques is presented. Then, existing algorithms to compress LF images are discussed in relation to the requirements of previously defined use cases. Finally, currently available objective and subjective evaluation methodologies to assess the quality of LF images are presented.

A. Light Field Image Acquisition

A digital 4D LF, which is a collection of perspective images, can be obtained by sampling the 4D LF function defined in (2). The density of the sampling depends on the acquisition technology used to capture the LF image. In general, depending on the requirements for baseline, different acquisition techniques can be used to capture LF images. More specifically, for baseline in a range of meters, one way of acquiring LF images is by means of a moving camera. The idea behind is that by moving a single camera throughout the scene, we can acquire the different perspective images that form our 4D LF. In this case, the sampling in xy plane depends on the camera resolution, and sampling in uv plane depends on the position of the capturing device and its shutter speed. Examples of such acquisition devices are the Stanford Spherical Gantry, a motorized gantry with four degrees of freedom that can be used to capture 4D LF [7] and Apple's setup to construct 360-degree cylindrical panoramic images [8]. 4D LF can also be acquired by using hand-held cameras, as long as their position on the uv plane can be precisely estimated [9].

Another approach is to construct an array of cameras with synchronized shutter speed capturing the 4D LF at once. In this case, the uv sampling depends on baseline parameter of the camera array grid. Using a camera array, a full 4D LF is formed and new views corresponding to narrower baseline parameter must be further synthesized if needed. An example of such acquisition technology is the Stanford Multi-Camera Array [10].

LF images can also be acquired from multi-view plus depth data [11]. In this case, the baseline can be wide or narrow, depending on how the data was created [12].

For LF image acquisition with narrow baseline a hand-held plenoptic camera capturing so called "single lens stereo" can be achieved [13]. It exploits an array of micro lenses placed in front of the sensor plane [14]. The aperture size of the main lens limits the possible viewpoints. The spatial resolution (xy plane) depends on the number of micro lenses, whereas the angular resolution (uv plane) depends on the number of pixels behind each micro lens. Hand-held cameras implementing this model were presented in [15] and are already widely available to consumers.[1,2] The raw image obtained with this type of cameras closely resembles the honeycomb array of lenses that has been used for the acquisition, and will be from now on referred to as a lenslet image. It is possible to convert the lenslet image to a 4D LF that effectively constitutes a sampling of (2).

B. Light Field Image Compression

Acquiring or synthesizing LF images creates a vast amount of data (around 150 MB for lenslet images with 15×15 viewpoints of resolution of 635×434, around 6.8 GB for 15×15 4K images acquired with a multi-camera array). Thus, a lot of research has been recently focused on finding efficient compression algorithms to effectively store and transmit LF images. Depending on the acquisition process, several approaches have been proposed. Early work focused on compressing synthetic 4D LF using disparity compensation [16]–[18] and geometry estimation [19], and enabling random access coding [3].

More recently, the effort has been focused on compression of LF images acquired through hand-held devices. Several compression algorithms have been proposed to directly compress lenslet images through intra coding, exploiting redundancies in its structure. For example, Perra proposed a lossless compression scheme based on adaptive prediction [20]. Li *et al.* incorporated a full inter prediction scheme in HEVC intra prediction, explicitly exploiting the redundancy in lenslet images [21], as well as using the disparity compensation and impainting to efficiently code lenslet images [22]. More recently, five proposals for lenslet image compression have been collected within the ICME 2016 Grand Challenge. Three proposals relied on improving HEVC intra compression efficiency by exploiting the redundancies in the lenslet image [23]–[25], whereas the rest used pseudo-temporal sequences to code the lenslet images [26], [27].

Another approach is based on compression of 4D LF, which can be created from raw lenslet image data using specific transformations. A precursor of the approach is proposed by Olsson *et al.* [28]. They propose the creation of sub-images from integral images. Such sub-images are then encoded through a pseudo-sequence using H.264. Choudhury *et al.* proposed to adapt the method of coded snapshots to LF image compression through random codes [29]. Dai *et al.* coded sub-aperture images using different scanning methods, including line and rotating scanning [30]. Helin *et al.* proposed predictive coding for sub-aperture views to achieve lossless compression [31].

No standard approach has been agreed on to compress LF images. Nevertheless, to help finding a standard representation for LF images, the JPEG standardization committee launched in 2014 a new activity called JPEG Pleno. Its goal is to create a standard framework for efficient storage and transmission of not only LF images, but also point-cloud, holographic and other plenoptic content. In particular, JPEG Pleno aims at finding an efficient way to represent plenoptic content, while, where needed, also offering compatibility with existing solutions, such as JPEG and JPEG 2000. In such a framework, it is essential to analyse and compare different approaches for compression of LF images as one of the plenoptic content variations, especially in relation to different use cases, in order to provide a solid guideline towards the creation of a standard method. Additionally, a call for proposals for both lenslet and high density camera array for 4D LF compression, aiming at the definition of a standard for compression of plenoptic content, has been issued during the 73^{rd} JPEG Meeting [32].

C. Light Field Image Quality Evaluation

Evaluation of visual quality is essential to design and improvement of coding solutions for LF images. Several publications have been devoted to comparison and evaluation of state of the art standard solutions. Alves *et al.* assessed the performance of existing still image coding solutions, such as JPEG 2000 and AVC, on lenslet images [33]. The objective evaluation was carried out using PSNR as a full reference metric. Similarly, Vieira *et al.* compared five different HEVC compatible coding of lenslet images with different data formats [34], again using PSNR as a full reference metric. Rizkallah *et al.* reported the impact of compression of LF images on refocusing and extended focus images through objective metrics [35].

A Grand Challenge was organized at ICME 2016 under a collaboration between Qualinet and JPEG standardization committee. The goal was to collect new compression solutions for LF images, and to evaluate them using both objective and subjective quality assessment methodologies [4]. The Grand Challenge, however, had some limitations. For starters, it required to have lenslet images in YUV 420 format as input of the compression, as well as output for the decompression. Moreover, the reference was obtained by performing the transformation to 4D LF after chroma subsampling was applied on the raw lenslet image, rather than transforming directly the lenslet image in RGB 444. The subjective assessments were carried out on five discrete views from each content, which were evaluated separately. Since the assessment was conducted separately on predefined

[1]https://www.raytrix.de/
[2]https://www.lytro.com/

views, it did not address the issue of evaluating global quality of experience offered by a compressed LF image.

Recently, we proposed a new methodology to evaluate a plenoptic content in an interactive way [5] allowing users to interact with LF images, visualize different views, apply refocusing, and globally evaluate the quality of LF images. To best of our knowledge, no extensive analysis has been carried out to assess which coding approach for LF image compression yields better results in terms of visual quality.

III. Light Field Coding Strategies

This section describes in details the coding approaches investigated in the evaluation process, including a thorough description of five selected algorithms to compress LF images.

Two main coding approaches can be considered for compression of LF images. Referring to the general diagram of workflow presented in Fig. 1, we can compress the data at two different stages. Compression can be performed on the raw sensor data that has been captured with the selected acquisition technology, after minimal processing, such as demosaicing and devignetting (point A in Fig. 1). The 4D LF can be recovered from the decompressed bitstream through extensive post-processing, involving camera and color calibration metadata which needs to be sent along with the bitstream. The second coding approach performs compression on the 4D LF obtained from the raw data (point B in Fig. 1). The 4D LF is a collection of perspective views which can be visualized as they are, or combined to create new interpolated views, synthetic aperture, refocusing, and extended focus. Since the transformation of raw sensor image data to 4D LF is performed before the compression step, no metadata is required for visualization. The compression solutions used to code the raw sensor data, as well as the transformation to 4D LF from the raw sensor data, strongly depend on the acquisition technology used to capture LF images. If compression is applied at point A, the selected scheme will profoundly differ based on the acquisition technology. On the other hand, a compression scheme operating at point B can compress 4D LF image information captured with any acquisition technology.

In order to compare the two coding approaches on a common ground, we decided to focus our attention on evaluating coding strategies for lenslet-based acquisition. Lenslet-based acquisition allows to compare the two approaches on the same image content captured within the same conditions. In this case, the raw sensor data is minimally pre-processed to obtain a lenslet image. From the lenslet image, the 4D LF can be recovered through rectification, calibration and extraction of perspective images, using camera and color calibration data. The extraction of perspective images from the lenslet image generates $N \times M$ views, depending on the uv resolution. However, the most external views contain too many distortions to be properly visualized. The 4D LF coding approach can take advantage of this fact by not coding the most distorted views, which will likely not be used in the visualization process, thus further reducing the size of the bitstream.

The rest of the section is organized as follows. The first coding approach, which deals with compression of lenslet images, is described. Then, the second coding approach, which focuses on compression of 4D LF obtained from lenslet images, is presented. Finally, one hybrid approach to compress lenslet images through transformation to 4D LF, introduced in ICME 2016 Grand Challenge, is detailed. Authors are aware of practical drawbacks and flaws in this solution. However, it was decided to include it in the evaluation process because of its optimal performance within the Grand Challenge, and because it represents a transition point between the two coding approaches. A summary of the compression schemes can be found in Table I.

A. Lenslet Image Compression

The lenslet coding approach performs compression on the lenslet image, obtained from the raw sensor data after demosaicing and devignetting. Fig. 2 depicts the workflow for the coding approach. The workflow was adopted following the definition of the ICME 2016 Grand Challenge, which required to perform compression on YUV 420 lenslet images in 8-bit precision.

The raw sensor data is first demosaiced, devignetted and clipped to 8 bits to obtain a lenslet image. The lenslet image is subsequently converted to YUV 420 format, and compressed and decompressed using the selected compression scheme. The output of the decompression step is then upsampled and converted to RGB 444. Conversion to RGB 444 format is required to perform the transformation from lenslet image to 4D LF. The 4D LF is created from the decompressed lenslet image using camera metadata. Color and gamma corrections are applied separately on each view. The perspective views forming the 4D LF can subsequentially be visualized on commercially available displays, or combined to create new interpolated views, synthetic aperture or refocusing effect.

Two compression algorithms ($P01$ and $P02$) compliant with the workflow depicted in Fig. 2 were evaluated. One compression scheme was selected among the best performing submitted to ICME 2016 Grand Challenge, and it was compared to an HEVC anchor. The anchor $P01$ exploits HEVC intra profile with default settings to compress the YUV 420 lenslet image. To perform the compression, the software x265 was used.[3] The second algorithm $P02$ uses a modified version of HEVC intra profile, which integrates Locally Linear Embedding (LLE) and Self Similarity (SS) to exploit the redundancies in the lenslet structure [25]. The image is partitioned into blocks using HEVC intra prediction scheme. LLE estimates the current block by selecting the best linear combination of k nearest neighbors through a least-square optimization problem. SS predicts the current block using the best among two blocks, one given by best block matching in the search window, the other chosen by searching for best linear combination between the first selected block and another block in the search window. The codec performs both LLE and SS, and then choses the prediction method that gives the smallest rate distortion cost.

[3]https://www.videolan.org/developers/x265.html

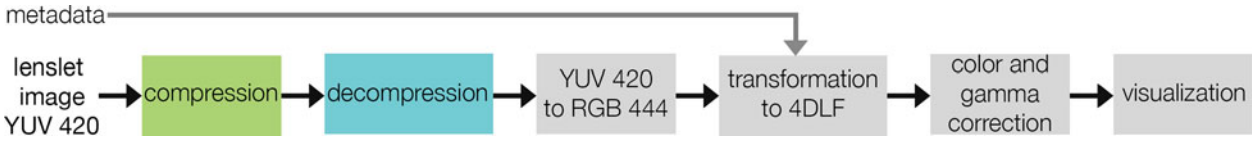


Fig. 2. Processing chain for lenslet image compression used for two compression algorithms (anchor $P01$, proponent $P02$).

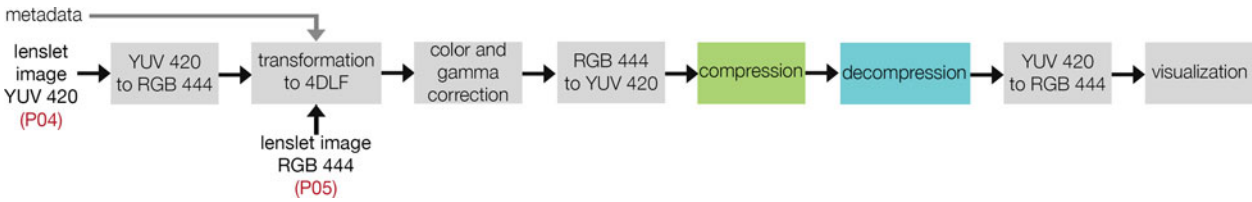


Fig. 3. Processing chain for 4D LF compression used for two compression algorithms (anchors $P04$ and $P05$).

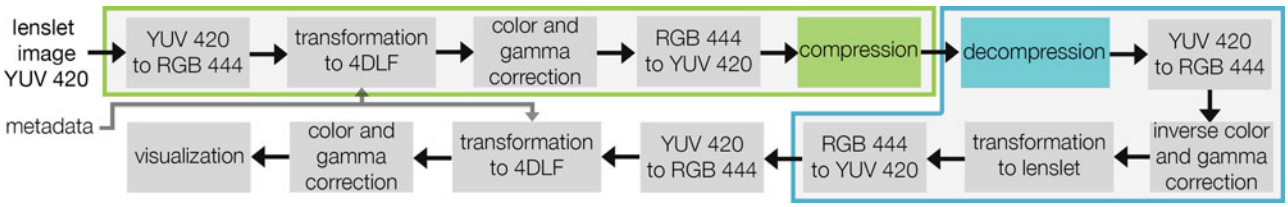


Fig. 4. Processing chain for hybrid compression of lenslet using intermediate 4D LF transformation (proponent $P03$). The green and blue blocks highlight how the compression step involves intermediate transformation to 4D LF, and the decompression step involves the inverse transformation to lenslet image.

B. 4D Light Field Compression

The 4D LF coding approach performs the compression on the 4D LF, obtained from the lenslet image, after color and gamma corrections. Fig. 3 depicts the workflow for this coding approach. Two anchors were created to assess the visual quality of this coding approach. The first anchor $P04$ assumes the same input as the compression schemes using the lenslet coding approach (YUV 420 lenslet images). The color space is then upsampled and converted again to RGB 444, to be used in the transformation process. To assess the effect of chroma subsampling of the lenslet image on the resulting quality of the final 4D LF, the second anchor $P05$ performs the compression on the 4D LF, obtained from lenslet in RGB 444 format, after color and gamma corrections (Fig. 3). In this case, the lenslet image is not transformed to YUV from RGB, and the color space is not subsampled before the transformation.

For both anchors, the 4D LF is created from the uncompressed lenslet image using camera metadata, and color and gamma corrections are applied separately on each view, prior to compression. Each view is converted from RGB 444 to YUV 420. The views are arranged in a pseudo-temporal sequence in spiral order, as depicted in Fig. 6. Due to the geometrical distortions present in the most external views, only a subset of the views is coded. Specifically, only the 13×13 internal views out of 15×15 views are encoded. The pseudo-temporal sequence is coded with HEVC software x265. In the decompression step, the views which have not been coded are replaced with copies of neighboring views, to reconstruct the 15×15 images that compose the 4D LF. After decompression, the views are upsampled and converted to RGB 444 and rearranged in the 4D LF. The perspective images composing the 4D LF can then be visualized on commercially available displays, or combined to create synthetic aperture, refocusing and new interpolated views.

C. Hybrid Compression of Lenslet Images

Among the participants to the ICME 2016 Grand Challenge, which required to have YUV 420 lenslet image as input and output of the compression and decompression step, one algorithm performed compression on lenslet images using intermediate transformation to 4D LF [26]. Fig. 4 depicts the workflow for this algorithm.

The algorithm $P03$ proposes a compression of 4D LF images based on pseudo-sequences of perspective views. Due to the constraints of the Grand Challenge, the YUV 420 lenslet image is first converted to RGB 444 color space, to be used in the transformation step. Then the lenslet is processed to obtain the perspective views that compose the 4D LF. The views are color and gamma corrected and then converted back to YUV 420. A subset of them is then rearranged in a specific coding order, that accounts for similarities between adjacent views, and coded using the JEM encoder.[4] In the decompression step, the views are rearranged in the 4D LF. Inverse color and gamma corrections are applied and the lenslet image is formed following the inverse process of the transformation to 4D LF.

The conversion from lenslet to 4D LF and back was needed to be compliant with the requirements of the grand challenge. However, it can be clearly seen that the proposed approach is

[4]https://jvet.hhi.fraunhofer.de/svn/svn_HMJEMSoftware/tags/HM-16.6-JEM-2.0rc1/

TABLE I
SUMMARY OF COMPRESSION SCHEMES

Proponents	Description
P01	Anchor: Lenslet image compressed using HEVC intra (default settings of x265 software implementation).
P02	Lenslet image compressed using HEVC intra with LLE and SS (software HM-14.0) [25].
P03	Hybrid approach: lenslet image compressed using intermediate transformation to 4D LF and HEVC (software JEM 2.0) [26].
P04	Anchor: 4D LF, obtained after chroma subsampling, compressed using HEVC (default settings of x265 software implementation).
P05	Anchor: 4D LF compressed using HEVC (default settings of x265 software implementation).



Fig. 5. Central viewpoint image from each content used in our experiment. (a) Bikes. (b) Stone_Pillars_Outside. (c) Fountain_&_Vincent_2. (d) Friends_1.

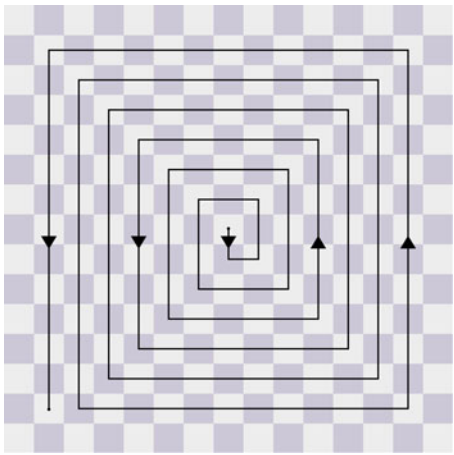


Fig. 6. Ordering of the views for coding.

hybrid, in the sense that it compresses lenslet images through transformation to 4D LF. The tranformation from lenslet images to 4D LF and back is lossless, as it is defined in [26].

IV. LIGHT FIELD QUALITY ASSESSMENT EXPERIMENT

This section describes the evaluation process in details. First, the data preparation process is presented, along with the coding conditions. Methodologies and metrics for objective and subjective evaluation are then presented in details.

A. Dataset Preparation and Coding Conditions

Four LF images, acquired by a Lytro Illum camera, were selected from the publicly available EPFL LF image dataset [36]. More specifically, *Bikes*, *Stone_Pillars_Outside*, *Fountain_&_Vincent_2* and *Friends_1* contents were selected for our experiments. The central view of each content used is depicted in Fig. 5. The images were carefully selected from those used in the ICME 2016 Grand Challenge [4] in order to provide a wide range of scenarios, containing details that would prove challenging for the compression algorithms. To obtain the 4D LF, the lenslet images were processed using the LF MATLAB toolbox [37], [38].

The compression algorithms were evaluated on four bitrates (corresponding to four compression ratios), namely $R1 = 1$ bpp $(10:1)$, $R2 = 0.5$ bpp $(20:1)$, $R3 = 0.25$ bpp $(40:1)$, $R4 = 0.1$ bpp $(100:1)$. The compression ratios are computed as ratios between the size of the uncompressed raw images in 10bit precision ($5368 \times 7728 \times 10$ bits $= 414\,839\,040$ bits) and the size of the compressed bitstreams.

The uncompressed reference was obtained by demosaicing, devignetting and clipping to 8 bits the raw sensor data, transforming it to 4D LF and applying color and gamma corrections. Unlike the reference used in ICME 2016 Grand Challenge, which used as a reference the 4D LF obtained from YUV 420 lenslet image, we obtain our reference from the lenslet image in RGB 444, without any chroma subsampling. This reference was selected to have a proper comparison with acquisition data obtained with minimal pre-processing. For this reason, chroma subsampling was not applied on the reference, since it alters the data.

A total of five compression schemes were evaluated. Each compression scheme was given a label, as stated before, for easier identification. A summary of the compression schemes can be found in Table I. It should be noted that the Quantization Parameters (QP) were selected to match the bitrates described above.

B. Objective Evaluation

To analyze the performance of evaluated coding schemes, PSNR was selected as a full reference metric. PSNR values were computed with respect to the uncompressed reference. The computation is thus performed on the 4D LF after color and gamma corrections (point B in Fig. 1). The PSNR metric was

TABLE II
VALUES OF REFOCUSING SLOPE FOR EACH CONTENT

	Slopes										
Content	1	2	3	4	5	6	7	8	9	10	11
Bikes	−10	−8	−6	−4	−2	0	2	4	6	8	10
Stone_Pillars_Outside	−10	−8	−6	−4	−2	0	2	4	6	8	10
Fountain_&_Vincent_2	−10	−8	−6	−4	−2	0	2	4	6	8	10
Friends_1	−5	−4	−3	−2	−1	0	1	2	3	4	5

adapted to better suit properties of LF images. Therefore, the PSNR value is computed on the Y channel as follows:

$$PSNR_Y(k,l) = 10\log_{10}\frac{255^2}{MSE(k,l)}, \tag{3}$$

in which k and l are the indexes of the acquired views. The $MSE(k,l)$ for each image is computed as follows:

$$MSE(k,l) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}[I(i,j) - R(i,j)]^2, \tag{4}$$

where m and n are the dimensions of one viewpoint image (i.e., $n = 625$, $m = 434$). $I(i,j)$ is the Y value for the selected acquired view in the evaluated 4D LF, whereas $R(i,j)$ is the corresponding value in the reference 4D LF. In the same way, the PSNR for the other two channels U and V is obtained. A weighted average [39] is then computed as follows:

$$PSNR_{YUV}(k,l) = \frac{6PSNR_Y(k,l) + PSNR_U(k,l) + PSNR_V(k,l)}{8} \tag{5}$$

The mean of all viewpoint images is subsequentially computed to have an average value for PSNR for Y channel and for YUV:

$$PSNR_{X_{mean}} = \frac{1}{(K-2)(L-2)}\sum_{k=2}^{K-1}\sum_{l=2}^{L-1}PSNR_X(k,l), \tag{6}$$

in which $K = 15$ and $L = 15$ represent the number of perspective views, and $X = Y$ and $X = YUV$ for Y channel and for YUV channels, respectively.

C. Subjective Evaluation

1) Interactive Test Methodology: A recently introduced methodology for interactive evaluation of plenoptic content was selected to perform the first of the two visual assessments [5]. The methodology is based on Double Stimulus Impairment Scale (DSIS) [40].

The participants were asked to interact with the LF images and to rate the level of impairment of the test LF image with respect to the reference, on a scale from 1 (*Very annoying*) to 5 (*Imperceptible*). Each LF image was presented together with the uncompressed reference in a side-by-side fashion. The position of the reference was set to either left or right for each experiment, and the participants were informed about its location on the screen.

For each stimulus, the central viewpoint image from the 4D LF was displayed. By clicking inside the displayed image and dragging the mouse, the other viewpoints from the 4D LF were accessed and displayed. Each image was displayed in its native resolution of 625×434 pixels.

Eleven refocused images were created for each content, using a modified version of the toolbox function *LFFiltShiftSum*. The function shifts all the perspective views according to a parameter, called slope, and performs a sum of the shifted images to obtain a single image that is refocused on a specific plane, which depends on the value of the slope. The number of images to be shifted and consequently summed defines the depth of field. Summing all 15×15 images creates the smallest depth of field, in which only one specific plane in the image is in focus. On the other hand, taking just the central image, which is equivalent to summing just 1×1 images, brings all the objects in focus (largest depth of field). For our tests, it was decided to sum images from index 3 to index 13 (11×11 images) to have a larger depth of field that still shows an effects of refocusing. The values of the slopes are summarized in Table II. The refocused images were accessible through a slider shown at the bottom of each stimulus. The slopes were selected so as to assure gradual transition between refocusing on the foreground and on the background with respect to semantically relevant objects in each content.

Before the experiments, a training session was organized to allow participants to get familiar with artifacts and distorsions in the test images. Five training samples were manually selected by expert viewers. In order not to influence the results, the training samples were created by compressing other content on various bitrates. The content used for the training was chosen from the same LF image database used for the test images [36]. The training samples were presented along with the uncompressed reference, exactly as they were shown in the test.

The experiment was split into two sessions. In each session, 40 stimuli were shown side by side with the uncompressed reference, corresponding to approximately 20 minutes per session. The display order of the stimuli was randomized, and the same content was never displayed twice in a row. Each subject took part in all the sessions, thus evaluating the entire set of stimuli. A break of ten minutes was enforced between the sessions to avoid fatigue. Before the test, one dummy sample was inserted to ease the participants into the task. The resulting scores from dummy stimuli were not included in the results.

A total of 24 subjects (19 males and 5 females) participated in the experiment, for a total of 24 scores per stimulus. Subjects

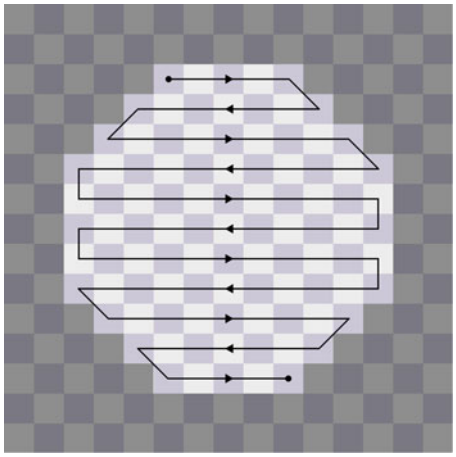


Fig. 7. Ordering of the views for animation for passive methodology.

were between 18 and 35 years old, with an average of 24.8 and a median of 25 years of age. All subjects were screened for correct visual acuity with Snellen charts, and color vision using Ishihara charts.

2) Passive Test Methodology: The second of the two visual assessments of quality was performed using a passive methodology, to ensure that all participants would visualize and rate exactly the same viewpoints and refocused views. The methodology is based on DSIS.

The participants were shown the LF content as a video sequence navigating between the viewpoints and the refocused images. Each stimulus was displayed alongside with the uncompressed reference, in a side by side fashion, and the subjects knew in advance on which side of the screen the reference was displayed.

Due to distortions caused by the lenslet structure, several viewpoints presented artefacts independent from the coding procedure, and thus had to be discarded, in order not to negatively influence the results. Therefore, only a subset of 97 out of 225 viewpoints was chosen to be displayed. Ten viewpoints per second were displayed, to ensure a smooth transition of the different viewpoints. The viewpoints were accessed from top to bottom and from left to right and right to left in alternate order (Fig. 7). At the end of the viewpoint animation, the eleven refocused images were displayed with a framerate of four refocused images per second, going from foreground to background and from background to foreground. The animation setup was chosen and validated by expert viewers in order to mimic the parallax effect, as well as to mimic the refocusing effect that occurs when trying to change the focal point. The total length of the animation for each stimulus was 14 seconds.

Test subjects were asked to rate the level of impairment of the test stimuli when compared to the uncompressed references. The rating was performed on a scale from 1 (Very annoying) to 5 (Imperceptible). Before the experiment, a training session was organized to allow participants to get familiar with artefacts and distorsions in the test images. Five training samples were manually selected by expert viewers. To help subjects localize and identify compression artefacts in the fast-paced video, the same content used in the test was selected for the training. The training samples were presented along with the uncompressed reference, exactly as they were shown in the test.

The experiment was split into two sessions. In each session, 40 stimuli were shown side by side with the uncompressed reference, corresponding to approximately 20 minutes per session. The display order of the stimuli was randomized, and the same content was never displayed twice in a row. Each subject took part in all the sessions, thus evaluating the entire set of stimuli. A break of ten minutes was enforced between the sessions to avoid fatigue.

A total of 29 subjects (24 males and 5 females) participated in the experiment, for a total of 29 scores per stimulus. Subjects were between 18 and 35 years old, with an average and median of 23 years of age.

3) Interactive Test Environment: To avoid the involuntary influence of external factors and to ensure the reproducibility of results, the laboratory for subjective video quality assessment was set up according to ITU recommendation BT.500-13 [40]. Professional Eizo ColorEdge CG301W 30-inch monitors with native resolution of 2560×1600 pixels were used for the test. The background color of the display was set to mid grey, according to requirements in ITU Recommendation ITU-R BT.2022 [41]. The monitors were calibrated using an i1Display Pro color calibration device according to the following profile: sRGB Gamut, D65 white point, 120 cd/m^2 brightness, and minimum black level of 0.2 cd/m^2. The room was equipped with a controlled lighting system that consisted of neon lamps with 6500 K color temperature, while the color of all the background walls and curtains present in the test area was mid grey. The illumination level measured on the screens was 15 lux. The distance of the subjects from the monitor was approximately equal to 7 times the height of the displayed content, conforming to requirements in ITU Recommendation ITU-R BT.2022 [41].

4) Passive Test Environment: To perform the tests, the QualityCrowd 2 framework [42] was used. Nonetheless, it should be noted that all the participants performed the tests in the same environment at the same time, with equal lighting conditions, using the same display model and the same screen resolution.

Since there is no browser video plugin capable of reliable real-time decoding and displaying for HEVC, the animations were encoded with AVC. A two-pass encoding was used and the deblocking filter was disabled to ensure transparency and to preserve the original blockiness artefacts when encoded at low bit rates. Expert viewing session conducted prior to the main subjective assessment concluded that the AVC video encoding was visually lossless. Selected settings for AVC coder are summarised in Table III.

5) Data Analysis: Outlier detection was performed according to the guidelines defined in ITU recommendation BT.500-13 [40]. One outlier was detected in both interactive and passive tests, and the relative scores were discarded, thus leading to 23 and 28 scores per stimulus, respectively. The Mean Opinion Score (MOS) was computed, separately for each methodology, for each coding condition j (i.e., each content, codec and compression ratio) as follows:

$$MOS_j = \frac{1}{N} \sum_{i=1}^{N} m_{ij}, \tag{7}$$

TABLE III
SELECTED SETTINGS FOR AVC CODER FOR PASSIVE METHODOLOGY

-r 30 -s <*size*> -f rawvideo -pix_fmt yuv420p -i <*input*> -c:v libx264 -profile:v high -x264opts no-scenecut:no-deblock:pass=1 -b:v 8M tmp.mp4
-r 30 -s <*size*> -f rawvideo -pix_fmt yuv420p -i <*input*> -c:v libx264 -profile:v high -x264opts no-scenecut:no-deblock:pass=2 -b:v 8M <*output*>

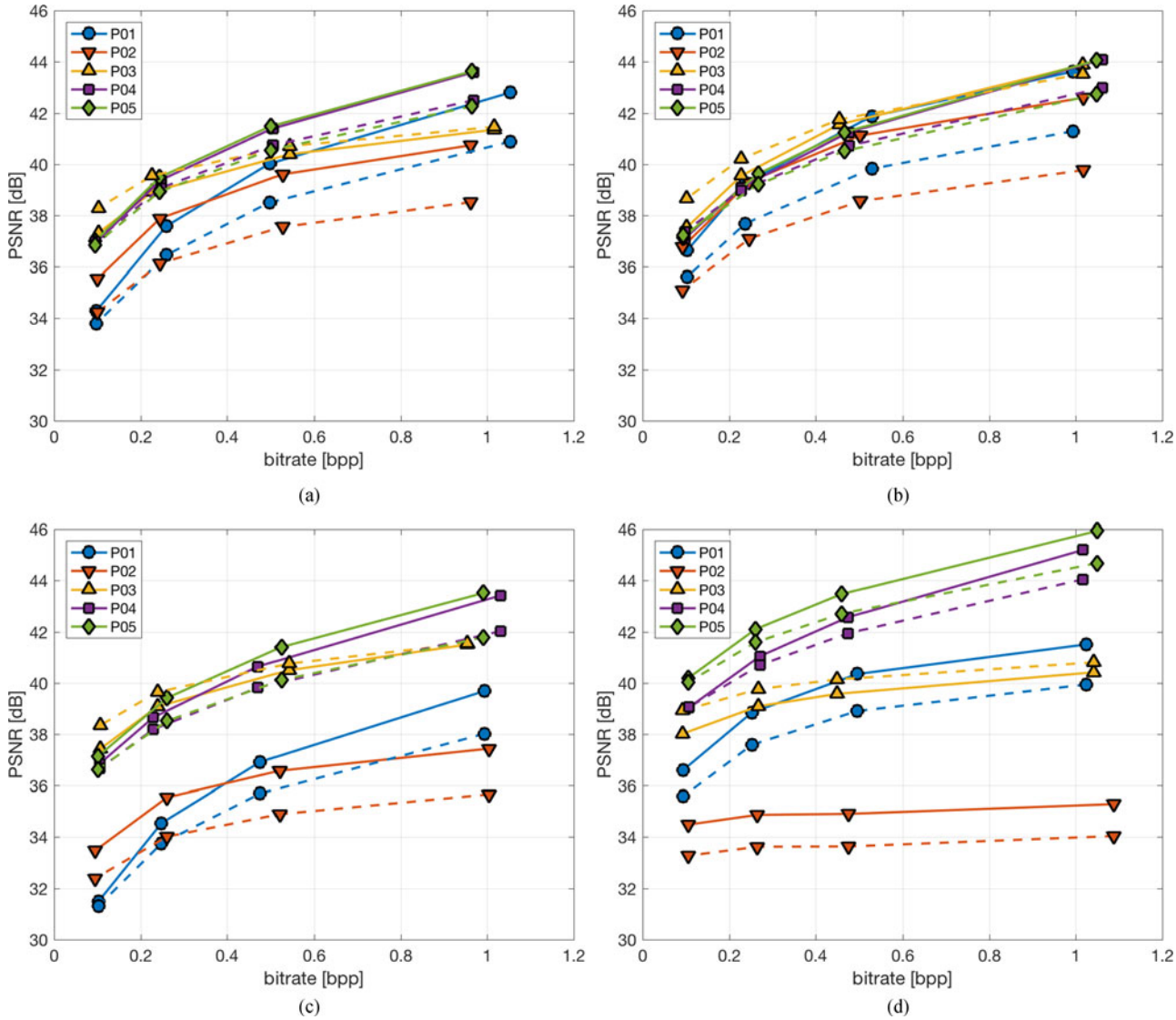


Fig. 8. Rate distortion plots for Y channel (solid line) and for YUV channels (dashed line). PSNR was computed on the 4D LF after color and gamma correction. (a) Bikes. (b) Stone_Pillars_Outside. (c) Fountain_&_Vincent_2. (d) Friends_1.

where N is the number of participants and m_{ij} is the score for stimulus j by participant i. The corresponding 95% confidence intervals were computed. To determine whether the results yield statistical significance, a one-sided Welch's test at 5% significance level was performed on the scores, with the following hypotheses:

$$H0 : MOS_A \leq MOS_B$$
$$H1 : MOS_A > MOS_B \,,$$

in which A and B are the codecs that are being compared. The test was performed for each compression ratio and for each content, separately for each methodology. If the null hypothesis were to be rejected, then it could be concluded that codec A performed better than codec B for the given content and compression ratio at a 5% significance level.

V. RESULTS AND DISCUSSION

In this section, results of the objective and subjective quality assessments are presented. Results on the coding approaches presented in Section III will be discussed separately. First, the lenslet image compression is analyzed. Then, the 4D LF compression is discussed. The hybrid approach is compared to the

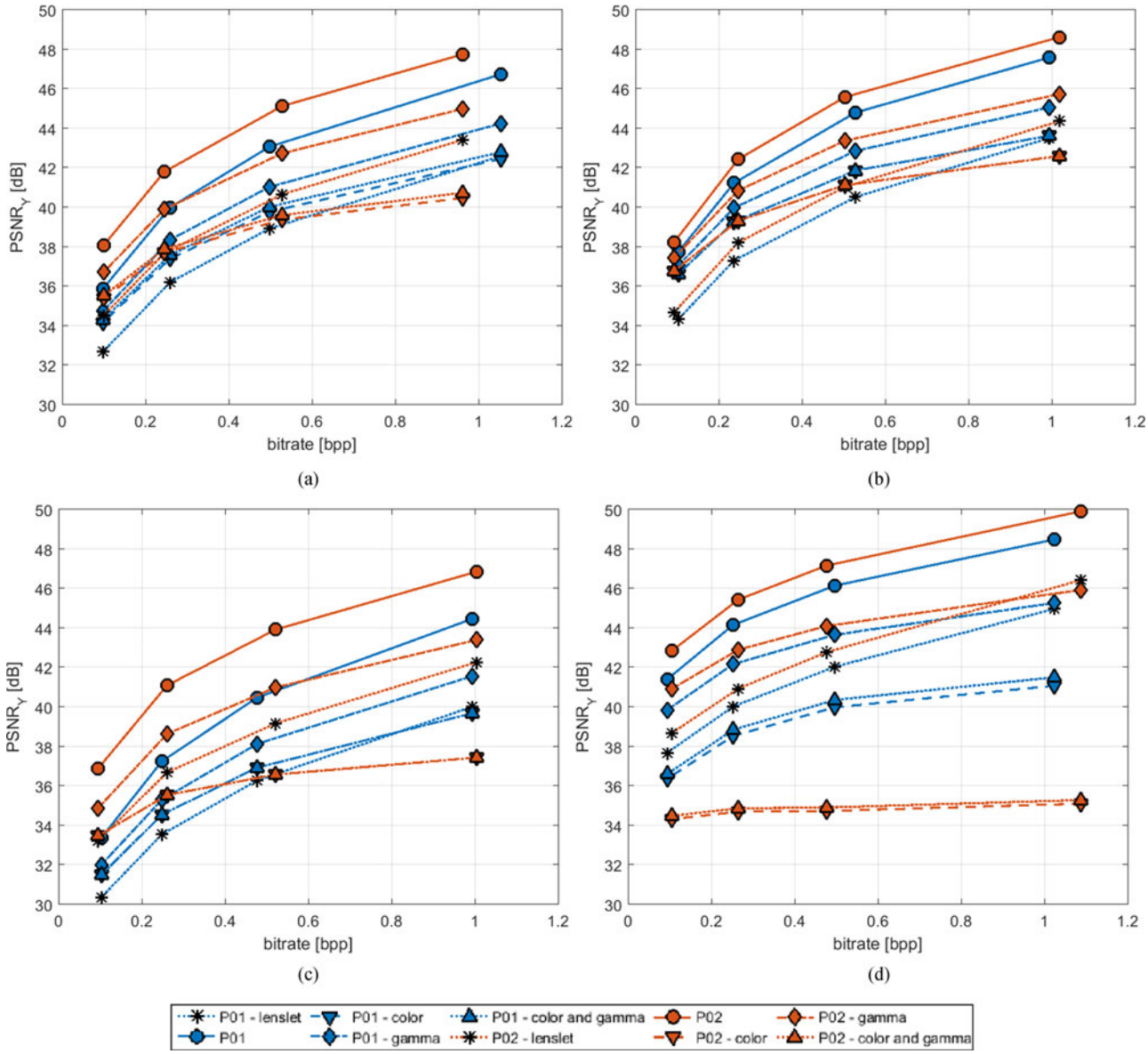


Fig. 9. Rate distorsion plots for Y channel. PSNR was computed at various stage of the pipeline (See Fig. 2). (a) Bikes. (b) Stone_Pillars_Outside. (c) Fountain_&_Vincent_2. (d) Friends_1.

other approaches. Finally, a comprehensive review of all the codecs is performed.

A. Compression of Lenslet Images

For PSNR computed on Y channel (Fig. 8, solid lines), the performance of the two codecs examined here ($P01$ and $P02$) strongly depends on the content, as it is common when computing PSNR. In general, $P01$ outperforms codec $P02$ for high bitrates. For low bitrates, $P02$ outperforms $P01$ for contents *Bikes* and *Fountain_&_Vincent_2*, and is outperformed in the remaining cases. PSNR computed on YUV channels (Fig. 8, dashed lines) shows similar trends.

Codec $P02$ has a particularly poor performance with content *Friends_1*, and in general performs worse than codec $P01$ for high bitrates. Results are particularly surprising since the codec proposed in $P02$ is supposed to improve the performance of HEVC Intra (anchor $P01$) with new prediction schemes. To better investigate the reasons behind this behaviour, we computed PSNR at different stages of the pipeline. Results from PSNR computation are shown in Fig. 9. In particular, we computed PSNR on the 4D LF without any color or gamma correction, on the color-corrected 4D LF, on the gamma-corrected 4D LF and when both corrections were applied on the 4D LF. Additionally, we compute PSNR on the lenslet image prior to the transformation, to better assess the performance of the two codecs on 2D images. The PSNR was computed with respect to the uncompressed reference at the same stage of the pipeline.

Results show that, on the lenslet image and on the 4D LF without any correction, $P02$ always outperforms $P01$. On the

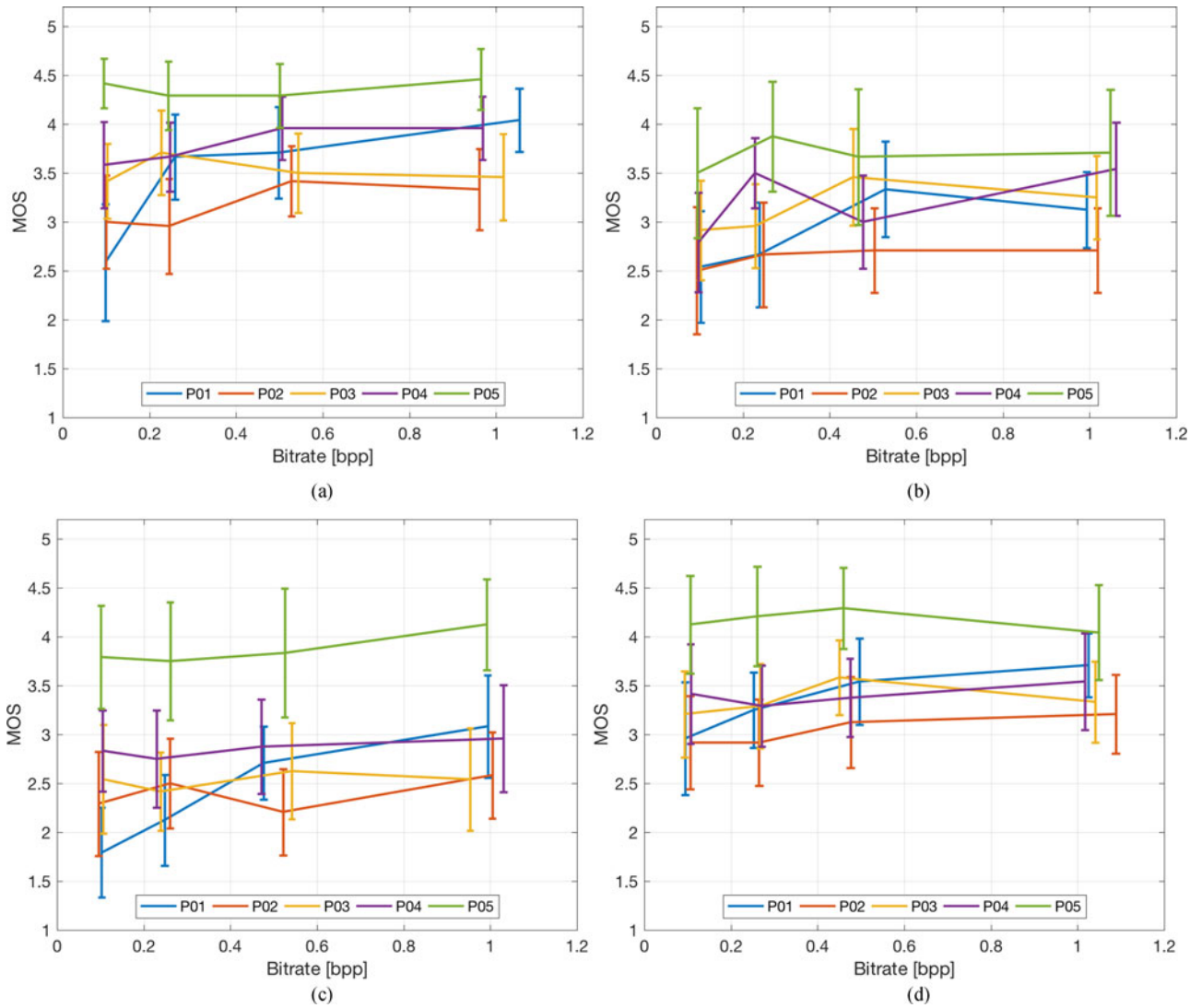


Fig. 10. Results of interactive subjective tests. MOS vs bitrate for all contents, with respective confidence intervals. (a) Bikes. (b) Stone_Pillars_Outside. (c) Fountain_&_Vincent_2. (d) Friends_1.

gamma-corrected 4D LF, $P02$ performs better than $P01$ on half of the contents. When color correction is applied on the 4D LF, however, we see a degradation in performance, with $P01$ outperforming $P02$ for high bitrates. This suggests that the prediction method, while working efficiently on compression of regular images, as proven by the results obtained on the lenslet image prior to transformation, adapts rather poorly to the peculiarities of LF images, and is more susceptible to errors after color correction is applied. Results from PSNR computed on YUV channels follow the same trend.

Results from both interactive and passive subjective evaluations show that $P01$ is performing significantly better than $P02$ for the highest bitrate. In particular, in the interactive test $P01$ is significantly better than $P02$ for all contents, whereas in the passive test it is significantly better for 3 out of 4 contents. For bitrate $= 0.5$ bpp, interactive tests show that $P01$ performs better than $P02$ for only 1 out of 4 contents, whereas passive tests indicate that it outperforms $P02$ on half of the contents. For lower bitrates (0.2 and 0.1 bpp) the difference between the two codecs is negligible.

B. Compression of 4D Light Field

As discussed in Section III, we want to analyse the effect of downsampling the lenslet image prior to transformation to 4D LF. For this reason, we compare the performance of $P04$, which uses a chroma subsampled version of the lenslet image, with $P05$, which creates the 4D LF from the lenslet image which has not been subsampled (see Fig. 3).

For PSNR computed on Y channel (Fig. 8, solid lines), the two codecs have similar performance for all bitrates for contents *Bikes* and *Stones_Pillars_Outside*, whereas for contents *Fountain_&_Vincent_2*, $P05$ performs better than $P04$ for all bitrates. Although downsampling of chroma channels should not affect the Y channel, color correction is applied on RGB channels of the single views, which are then converted to YUV

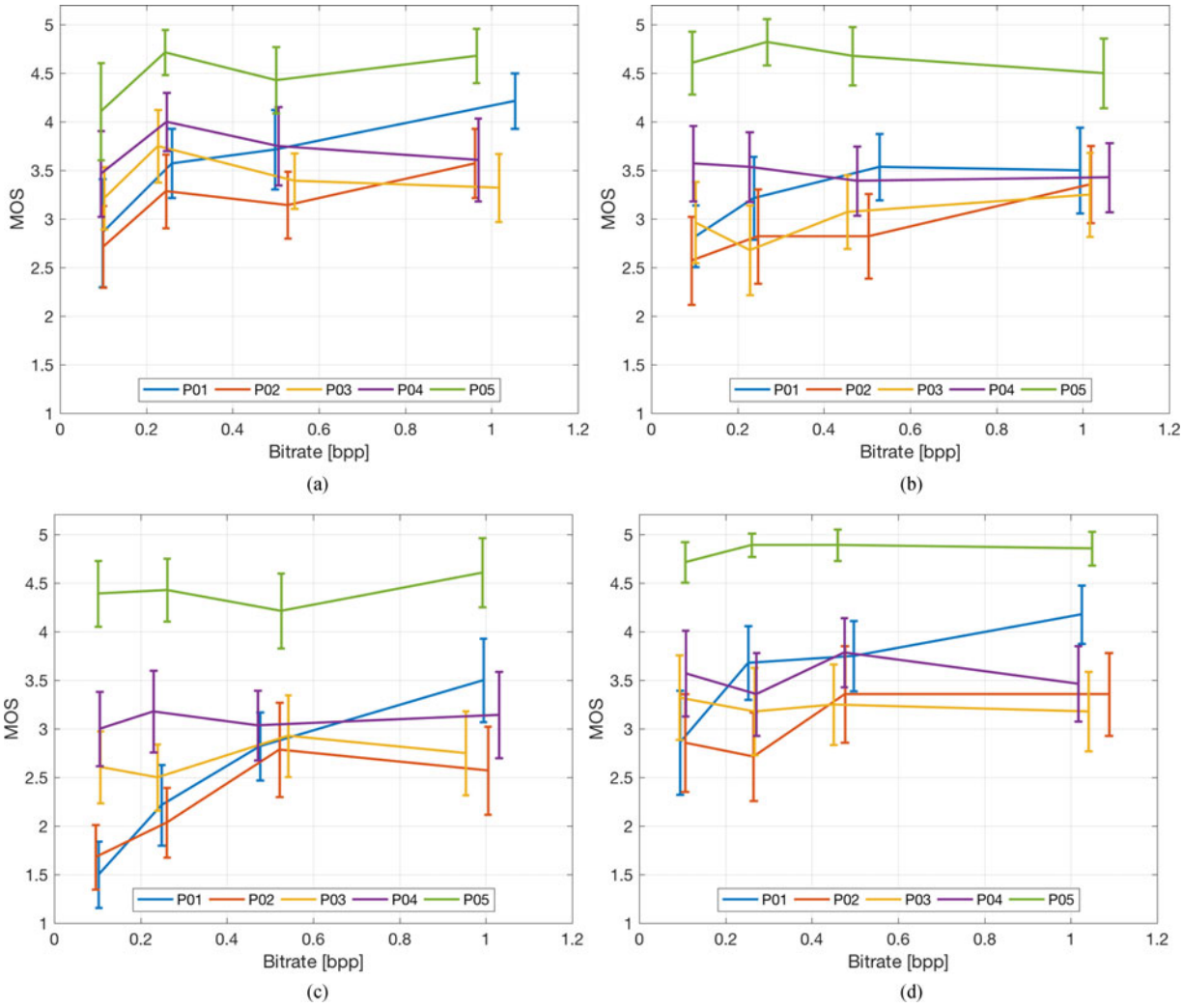


Fig. 11. Results of passive subjective tests. MOS vs bitrate for all contents, with respective confidence intervals. (a) Bikes. (b) Stone_Pillars_Outside. (c) Fountain_&_Vincent_2. (d) Friends_1.

to compute the PSNR. The downsampling thus affects the Y channel as well.

Results are similar for PSNR computed on YUV channels (Fig. 8, dashed lines), although for content *Fountain_&_Vincent_2*, the difference between $P04$ and $P05$ is now negligible.

Results from subjective evaluations and pairwise comparison (Figs. 10–13) show a stronger preference for codec $P05$ when compared to codec $P04$. In particular, results from the interactive tests show that, for the highest bitrate, $P05$ performs significantly better than $P04$ for two out of four contents. For bitrate $= 0.25$ bpp, $P05$ performs better on three out of four contents, whereas for the remaining two bitrates (0.5 bpp, 0.1 bpp) it always performs significantly better than $P04$.On the other hand, results from the passive tests show that $P05$ always performs significantly better than $P04$, for all bitrates.

C. Hybrid Compression of Lenslet Images

As seen in Section III, the third compression scheme $P03$ is compressing 4D LF and then converting them back to lenslet images. It is thus worthy of note to compare the performances of $P01$, $P02$ and $P03$, since they have the same input and output in the compression and decompression steps, although they use different approaches.

From the objective metric point of view, $P01$ and $P03$ outperform codec $P02$ for high bitrates. For low bitrates, $P03$ always outperforms $P01$ and $P02$, although in case of content *Stone_Pillars_Outside*, the difference between the codecs is negligible. For PSNR computed on YUV channels (Fig. 8, dashed lines), codec $P03$ outperforms the others for all contents. Interestingly enough, for codec $P03$ PSNR computed on YUV channels always has higher values than PSNR computed on the Y channel, while for all the other codecs the opposite is true. One possible explanation for this peculiar behavior is that the inverse color and gamma transformation applied before transforming the 4D LF back to lenslet has an effect on the final color performance, leading to better results in the YUV channels.

The subjective evaluation results do not show the same trends as the objective results (Figs. 10–13). In particular, results from the interactive tests show that for the lowest bitrate (0.1 bpp) $P03$ outperforms $P01$ on two out of four contents and never

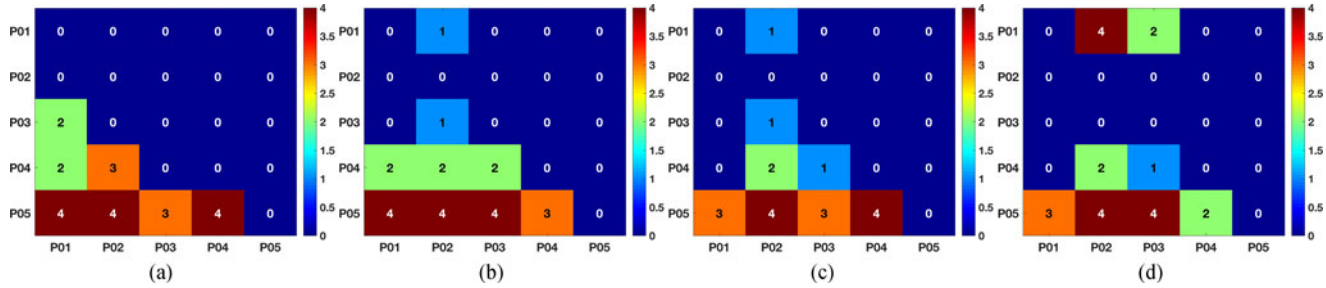


Fig. 12. Pairwise comparison results for interactive subjective tests. Each cell contains the number of contents for which the null hypothesis was rejected, for each compression ratio. The null hypothesis is defined as $MOS_i \leq MOS_j$, in which i indicates the row and j the column of the matrix. (a) R4 (b) R3 (c) R2 (d) R1.

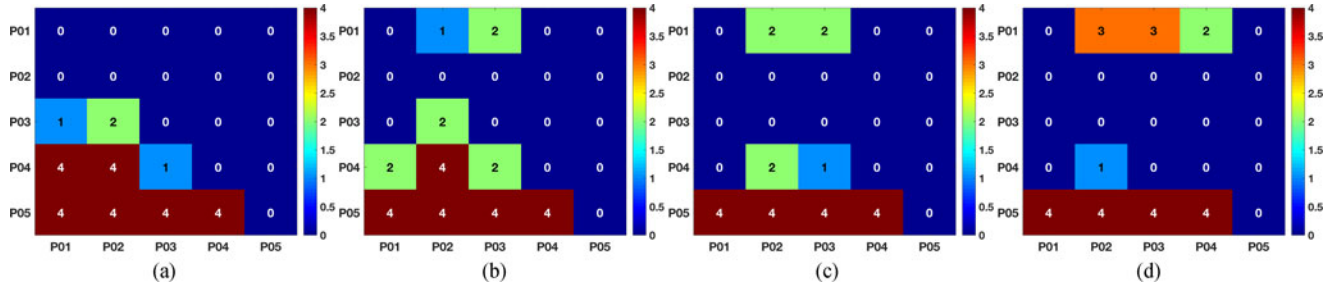


Fig. 13. Pairwise comparison results for passive subjective tests. Each cell contains the number of contents for which the null hypothesis was rejected, for each compression ratio. The null hypothesis is defined as $MOS_i \leq MOS_j$, in which i indicates the row and j the column of the matrix. (a) R4 (b) R3 (c) R2 (d) R1.

outperforms $P02$, whereas the passive tests show that $P03$ performs better than $P01$ on only 1 out of 4 contents, and performs better than $P02$ on 2 out of 4 (Fig. 13(a)). For intermediate bitrates (0.5 bpp and 0.25 bpp), interactive tests show that $P01$ and $P03$ both perform significantly better than $P02$ on one out of four contents, whereas passive tests additionally show that $P01$ performs significantly better than $P03$ on half of the contents for both bitrates. For the highest bitrate, $P01$ performs significantly better than $P03$ on at least half of the contents (3 out of 4 in case of passive tests, 2 out of 4 in case of interactive tests), and outperforms $P02$ in the majority of cases (3 out of 4 in case of passive tests, 4 out of 4 in case of interactive tests).

For the objective evaluation, the hybrid scheme $P03$ performs better than the other lenslet compression schemes. However, results from the subjective evaluation suggest that the difference in performance with respect to $P01$ (simple HEVC Intra) is negligible for low bitrates, and leads to poorer results for the highest bitrates.

Since $P03$ compresses lenslet images through transformation to 4D LF, it is useful to compare its performance to the performance of $P04$. For PSNR computed on Y channel (Fig. 8, solid lines), the performance of codecs $P03$ and $P04$ strongly depends on the content, as expected. For high bitrates, $P04$ performs better than $P03$, with the notable exception of content *Stones_Pillars_Outside*, in which codec $P03$ performs slightly better for all bitrates. For low bitrates, however, $P04$ performs slightly worse than $P03$ for all contents except *Friends_1*, for which $P04$ performs better than $P03$. PSNR computed on YUV channels (Fig. 8, dashed lines) show similar trends.

Results from subjective evaluation and pairwise comparison (Figs. 10–13), however, show that codec $P03$ never performs significantly better than codec $P04$. In particular, for the lowest bitrate, results from the interactive tests indicate that no codec performs significantly better than the other, whereas results from passive tests suggest that $P04$ performs better than $P03$ on 1 out of 4 contents. For bitrate $= 0.25$ bpp and 0.5 both interactive and passive tests agree that $P04$ outperforms $P03$ for 2 out of 4 contents and 1 out of 4 contents, respectively. For the highest bitrate, interactive tests indicate that $P04$ performs significantly better than $P03$ on 1 out of 4 contents, whereas for the passive tests they are statistically equivalent for all contents.

D. General Discussion

In general, both objective and subjective results show that coding 4D LF (point B in Fig. 1) leads to better performance when compared to coding lenslet images directly. In particular, pseudo-temporal ordering of 4D LF, obtained from RGB 444 lenslet image, performs significantly better than the other proposals for at least half of the contents for all bitrates examined in the subjective assessment of quality, showing that chroma subsampling of lenslet images can lead to a considerable reduction in visual quality. It is worth noting, however, that results from passive tests show that $P01$ performs statistically better than $P04$ on 2 out of 4 contents for the highest bitrate (Fig. 13(d)).

Comparison of different lenslet image compression algorithms shows that improvements in performance for 2D image coding do not necessarily result in better visual quality of LF image. In particular, whereas HEVC intra with LLE and SS has

better performance in objective evaluation carried out on lenslet images and 4D LF without color correction, it performs significantly worse when color correction is applied. Further work on lenslet image compression should address the effect of color correction on the final 4D LF, and propose new strategies to appropriately cope with this issue.

Coding 4D LF has the benefit of not requiring any metadata to be correctly displayed. Moreover, it can be used to code 4D LFs acquired with different acquisition technologies. Since the most distorted views in the 4D LF can be discarded in the compression process, it also allows for bitrate saving. As we previously mentioned, the transformation to 4D LF is an additional step that would be not suitable for low-memory devices. Thus, if consumers' market is the desired target, a solution that does not require any transformation would be preferrable. In this case, coding 4D LFs seems the most suitable choice.

The additional step of converting to 4D LF is not an issue if the target is the professional market. However, fidelity to acquisition parameters is of paramount importance. As seen before, chroma subsampling leads to poorer performances, especially after color correction has been applied. On the other hand, coding 4D LF leads to discarding metadata, which could be used in post-processing softwares, as well as potentially discarding some heavily distorted views. In this case, both approaches presented in this paper do not seem suitable. A new approach should be designed, aimed at high fidelity to acquisition parameters.

VI. Conclusion

In this paper, two different coding approaches for light field image compression were defined, described, and evaluated. Objective and subjective quality assessments of five different compression algorithms, following the aforementioned coding approaches, were conducted. Experimental results provide some insights on the impact of compression algorithms within, as well as across predefined use cases, on the perceived quality. This reveals the necessity of further investigations and improvements of compression algorithms especially in terms of processing of the metadata related to light field image data rendering.

More specifically, subjective quality evaluations show that one coding approach, namely, compressing 4D LF, yields significantly better results in terms of visual quality for all bitrates when compared to compressing lenslet images. The 4D LF coding approach is particularly suitable for general consumers' use case, since it does not involve additional computations at the decoder side to be properly rendered. Moreover, the coding approach does not require metadata to be successfully decoded and displayed, thus reducing the bitrate. Finding a successfull approach for the professional market, however, is still an open issue. A new method for compressing lenslet images while taking into account color fidelity must be designed for this type of market. Further research should focus on how to modify the proposed compression algorithm for light field images to further improve the performance and to meet the needs of all use cases.

References

[1] A. Gershun, "The light field," *Stud. Appl. Math.*, vol. 18, no. 1–4, pp. 51–151, 1939.

[2] E. H. Adelson and J. R. Bergen, *The Plenoptic Function and the Elements of Early Vision*. Cambridge, MA, USA: Vision and Modeling Group, Media Laboratory, Massachusetts Inst. Technol., 1991.

[3] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Annu. Conf. Comput. Graph. Interactive Tech.*, 1996, pp. 31–42.

[4] I. Viola, M. Rerabek, T. Bruylants, P. Schelkens, F. Pereira, and T. Ebrahimi, "Objective and subjective evaluation of light field image compression algorithms," in *Proc. 32nd Picture Coding Symp.*, 2016, pp. 1–5.

[5] I. Viola, M. Řeřábek, and T. Ebrahimi, "A new approach to subjectively assess quality of plenoptic content," in *Proc. SPIE Opt. Eng.+ Appl*. 2016, pp. 99 710X–99 710X.

[6] I. Viola, M. Rerabek, and T. Ebrahimi, "Impact of interactivity on the assessment of quality of experience for light field content," in *Proc. 9th Int. Conf. Quality Multimedia Exp.*, 2017, pp. 1–6.

[7] "The (new) stanford light field archive," Jul. 2016. [Online]. Available: http://lightfield.stanford.edu/papers.html

[8] S. E. Chen, "Quicktime VR: An image-based approach to virtual environment navigation," in *Proc. 22nd Annu. Conf. Comput. Graph. Interactive Tech.*, 1995, pp. 29–38.

[9] C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen, "Unstructured lumigraph rendering," in *Proc. 28th Annu. Conf. Comput. Graph. Interactive Techn*., 2001, pp. 425–432.

[10] B. Wilburn *et al.*, "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, 2005.

[11] A. Ouazan, P. T. Kovacs, T. Balogh, and A. Barsi, "Rendering multiview plus depth data on light-field displays," in *Proc. 3DTV Conf.: True Vision-Capture, Transmiss. Display 3D Video*, 2011, pp. 1–4.

[12] F. Zilly, C. Riechert, M. Müller, and P. Kauff, "Generation of multi-view video plus depth content using mixed narrow and wide baseline setup," in *Proc. 3DTV-Conf.: True Vision-Capture, Transmiss. Display 3D Video*, 2012, pp. 1–4.

[13] E. H. Adelson and J. Y. Wang, "Single lens stereo with a plenoptic camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 99–106, Feb. 1992.

[14] M. Levoy, "Light fields and computational imaging," *IEEE Comput.*, vol. 39, no. 8, pp. 46–55, Aug. 2006.

[15] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Comput. Sci. Tech. Report*, vol. 2, no. 11, pp. 1–11, 2005.

[16] M. Magnor and B. Girod, "Data compression for light-field rendering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 3, pp. 338–343, Apr. 2000.

[17] A. Jagmohan, A. Sehgal, and N. Ahuja, "Compression of lightfield rendered images using coset codes," in *Proc. Signals, Syst. Comput. Conf. Rec. 37th Asilomar Conf.*, 2003, vol. 1, pp. 830–834.

[18] B. Girod, C.-L. Chang, P. Ramanathan, and X. Zhu, "Light field compression using disparity-compensated lifting," in *Proc. Int. Conf. Multimedia Expo, ICME'03*, 2003, vol. 1, pp. I–373.

[19] X. Zhu, A. Aaron, and B. Girod, "Distributed compression for large camera arrays," in *Proc. IEEE Workshop Stat. Signal Processing*, 2003, pp. 30–33.

[20] C. Perra, "Lossless plenoptic image compression using adaptive block differential prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 1231–1234.

[21] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Efficient intra prediction scheme for light field image compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 539–543.

[22] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Scalable coding of plenoptic images by using a sparse set and disparities," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 80–91, Jan. 2016.

[23] Y. Li, R. Olsson, and M. Sjöström, "Compression of unfocused plenoptic images using a displacement intra prediction," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2016, pp. 1–4.

[24] C. Conti, P. Nunes, and L. D. Soares, "HEVC-based light field image coding with bi-predicted self-similarity compensation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2016, pp. 1–4.

[25] R. Monteiro *et al.*, "Light field HEVC-based image coding using locally linear embedding and self-similarity compensated prediction," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2016, pp. 1–4.

[26] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2016, pp. 1–4.

[27] C. Perra and P. Assuncao, "High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2016, pp. 1–4.
[28] R. Olsson, M. Sjostrom, and Y. Xu, "A combined pre-processing and H.264-compression scheme for 3D integral images," in *Proc. IEEE Int. Conf. Image Process*, 2006, pp. 513–516.
[29] C. Choudhury, Y. Tarun, A. Rajwade, and S. Chaudhuri, "Low bit-rate compression of video and light-field data using coded snapshots and learned dictionaries," in *Proc. IEEE 17th Int. Workshop Multimedia Signal Process.*, 2015, pp. 1–6.
[30] F. Dai, J. Zhang, Y. Ma, and Y. Zhang, "Lenselet image compression scheme based on subaperture images streaming," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 4733–4737.
[31] P. Helin, P. Astola, B. Rao, and I. Tabus, "Sparse modelling and predictive coding of subaperture images for lossless plenoptic image compression," in *Proc. 3DTV-Conf.: True Vision-Capture, Transmiss. Display 3D Video*, 2016, pp. 1–4.
[32] ISO/IEC JTC 1/SC29/WG1 JPEG, "Jpeg pleno call for proposals on light field coding," Doc. N73013, Chengdu, China, Oct. 2016.
[33] G. Alves, F. Pereira, and E. A. da Silva, "Light field imaging coding: Performance assessment methodology and standards benchmarking," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2016, pp. 1–6.
[34] A. Vieira, H. Duarte, C. Perra, L. Tavora, and P. Assuncao, "Data formats for high efficiency coding of Lytro-Illum light fields," in *Proc. Int. Conf. Image Process. Theory, Tools Appl.*, 2015, pp. 494–497.
[35] M. Rizkallah, T. Maugey, C. Yaacoub, and C. Guillemot, "Impact of light field compression on refocused and extended focus images," in *Proc. 24th Eur. Signal Process. Conf.*, 2016, pp. 898–902.
[36] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Proc. 8th Int. Conf. Quality Multimedia Exp.*, 2016, pp. 1–2.
[37] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1027–1034, Jun. 2013.
[38] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Linear volumetric focus for light field cameras," *ACM Trans. Graph.*, vol. 34, no. 2, pp. 1–20, Feb. 2015.
[39] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards—Including high efficiency video coding (HEVC)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1669–1684, 2012.
[40] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R BT.500-13, Jan. 2012.
[41] *General Viewing Conditions for Subjective Assessment of Quality of SDTV and HDTV Television Pictures on Flat Panel Displays*, ITU-R BT.2022, Aug. 2012.
[42] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Qualitycrowd—A framework for crowd-based quality evaluation," in *Picture Coding Symp.*, 2012, pp. 245–248.



Irene Viola received the B.Sc. degree in cinema and media engineering and the M.Sc. degree in computer engineering from the Polytechnic University of Turin, Turin, Italy, in 2013 and 2015, respectively. She is currently working toward the Ph.D. degree in the Multimedia Signal Processing Group (MMSPG), under the supervision of Professor Touradj Ebrahimi, at the Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland.

Her research interest includes image and video processing, compression, and evaluation, with a focus on light field coding.



Martin Řeřábek received the Ph.D. degree in electrical engineering from the Faculty of Electrical Engineering (FEE), Czech Technical University in Prague, Prague, Czech Republic, in 2013. He is a Postdoctoral Researcher at the Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland. He is the author or the coauthor of more than 70 research publications. His main research interest is focused on image and video processing and compression, evaluation and improvement of quality of experience in immersive multimedia technologies, biomedical, astronomical, and aerial image processing.



Touradj Ebrahimi (M'95) is currently a Professor at Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland, heading its Multimedia Signal Processing Group. He is also the Convener of JPEG standardization Committee. He was also adjunct Professor with the Center of Quantifiable Quality of Service at Norwegian University of Science and Technology from 2008 to 2012. His research interests include still, moving, and 3-D image processing and coding, visual information security (rights protection, watermarking, authentication, data integrity, steganography), new media, and human computer interfaces (smart vision, brain computer interface). He is the author or the coauthor of more than 200 research publications, and holds 14 patents.

Pseudo-Sequence-Based 2-D Hierarchical Coding Structure for Light-Field Image Compression

Li Li, *Member, IEEE*, Zhu Li, *Senior Member, IEEE*, Bin Li, *Member, IEEE*, Dong Liu, *Member, IEEE*, and Houqiang Li, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a pseudo-sequence-based two-dimensional (2-D) hierarchical coding structure for light-field image compression. In the proposed scheme, we first decompose the light-field image into multiple views and organize them into a 2-D coding structure according to the spatial coordinates of the corresponding microlens. Then, we mainly develop three algorithms to optimize the 2-D coding structure. First, we propose a 2-D hierarchical coding structure with a limited number of reference frames to exploit the intercorrelations among various views. To be more specific, we divide all the views into four quadrants, and all the views are encoded one quadrant after another to reduce the reference buffer size as much as possible. Inside each quadrant, all the views are encoded hierarchically in both horizontal and vertical directions to fully exploit the correlations among different views. Second, we propose to use the distance between the current view and its reference views instead of the picture order count difference as the criterion for selecting better reference frames for each inter view. The distance-based criterion is also applied to the motion vector scaling process to obtain more accurate motion vector predictors. Third, an optimal bit allocation algorithm taking the influence of the various views on the following encoding views into account is proposed to further exploit the intercorrelations among various views and improve coding efficiency. The entire scheme is implemented in the reference software of high efficiency video coding. The experimental results demonstrate that the proposed novel pseudo-sequence-based 2-D hierarchical structure can achieve maximum 28.4% bit-rate savings compared with the previous pseudo-sequence-based light-field image compression method.**

***Index Terms*—Bit allocation, hierarchical coding structure, high efficiency video coding, light-field image, pseudo sequence.**

Manuscript received December 28, 2016; revised April 16, 2017 and May 21, 2017; accepted June 26, 2017. Date of publication July 11, 2017; date of current version October 23, 2017. This work was supported in part by UMKC startup grant, in part by Natural Science Foundation of China under Contract 61325009, and in part by 973 Program under Contract 2015CB351803. The guest editor coordinating the review of this paper and approving it for publication was Dr. Yebin Liu. *(Corresponding author: Zhu Li.)*

L. Li is with the University of Missouri–Kansas City, Kansas City, MO 64110 USA, and also with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei 230027, China (e-mail: lil1@umkc.edu).

Z. Li is with the University of Missouri–Kansas City, Kansas City, MO 64110 USA (e-mail: lizhu@umkc.edu).

B. Li is with the Microsoft Asia, Beijing 100010, China (e-mail: libin@microsoft.com).

D. Liu and H. Li are with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei 230027, China (e-mail: dongeliu@ustc.edu.cn; lihq@ustc.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2725198

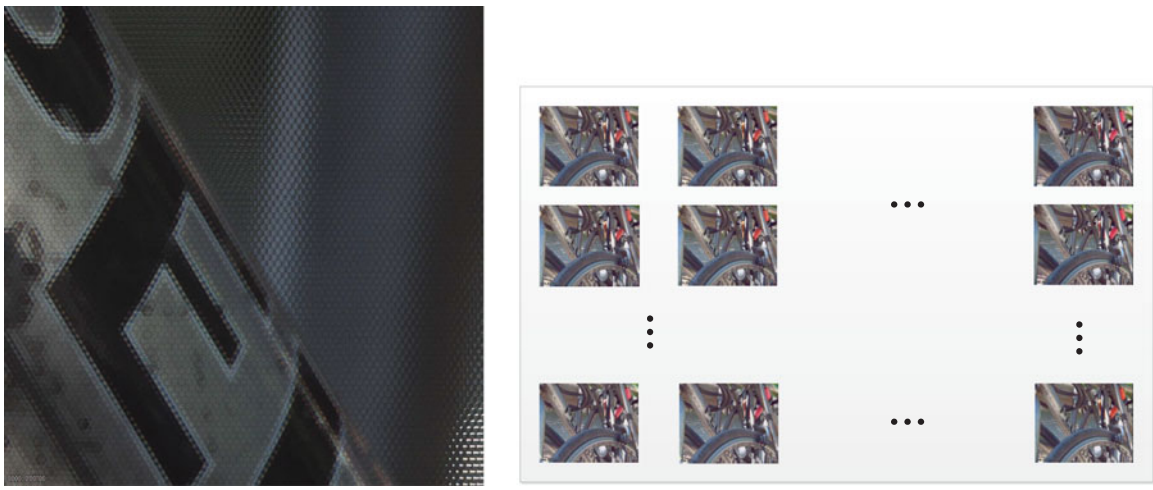
I. Introduction

THE light-field (LF) image [1], also known as the plenoptic image, contains the information about not only the intensity of light in a scene but also the direction of the light rays in space. The capture of the light or motion is becoming more and more advanced along with the development of the capturing methods [2]. Typically, the LF image is acquired with a plenoptic camera, which places an array of micro-lenses in front of a conventional image sensor. The light beams coming from the object with various angles are firstly refracted through the microlens array. Then they will be captured by the traditional 2-D image sensor to generate the raw sensor data of the LF image. In this way, each micro-lens works as an individual small low resolution image camera conveying a particular perspective of the object in slightly different angles. The raw sensor data will then be converted into the LF image after demosaicing, devignetting, and data structure conversion. As shown in Fig. 1, the LF Toolbox v0.4 [3] can convert the raw sensor data generated by the commercial LF camera Lytro Illum in Fig. 1(a) into the LF structure as shown in Fig. 1(b). In Fig. 1(b), each rectangle represents a 2-D view obtained from the contributions of all micro-lenses of the camera. From Fig. 1(b), we can see that the LF image is actually with 4-D information including not only the spatial information (similar to 2-D images) but also the angular information (different views).

Since the LF image records the light rays of a scene of interest, it can naturally provide the benefits of rendering new views not only for the changed viewpoint but also for the changed focal point. Recently, especially due to the emergence of the commercial LF cameras, the LF image is becoming a more and more attractive solution to 3-D imaging and sensing. However, the widely use of the LF image is still restricted by its massive size. Since the LF image is with 4-D information, even if the spatial resolution of one view is quite small, the raw data of a LF image with hundreds of views is still very large. For example, the resolution of a raw LF image generated by Lytro Illum [4] containing the captured field of light information is 7728×5368 pixels. Besides the huge image size, since the LF image is generated from a micro-lens array, its characteristic is entirely different from the general 2-D image as shown in Fig. 1(a), which makes it even harder to compress.

The recent researches for LF image compression can be mainly divided into two kinds both making full use of the classic image/video coding standards, such as JPEG [5], H.264/Advanced Video Coding [6], and H.265/High Efficiency

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.



(a) (b)

Fig. 1. Example of the raw sensor data and its corresponding LF image data. (a) Plenoptic camera raw sensor data after demosaicing and devignetting. (b) Light-field data structure.

Video Coding (HEVC) [7]. One kind of methods, called as the self-similarity based LF image compression [8], tries to compress the LF image using the commonly used image compression standards. Besides, since the different views of the LF image are quite similar to each other, the self-similarity compensated prediction and estimation (similar to intra block copy [9]) are usually employed to exploit the correlations among various views. However, the self-similarity based LF image compression methods are not flexible enough to fully exploit the correlations among various views.

The other kind of methods, named as the pseudo-sequence-based LF image compression [10], attempts to decompose the LF image into multiple views and tries to utilize the inter prediction in video coding standards to exploit the temporal correlations among various views. This kind of methods has the potential to achieve a very good coding efficiency since the inter prediction in video coding standards such as HEVC is quite efficient. However, since the LF image is very dense and with multiple views, the current works, which are all based on very coarse reference structures, are unable to make full use of the correlations among various views. How to organize the various views with an efficient coding order and a flexible reference structure remains a very serious problem.

Therefore, in this paper, we follow the approaches of the pseudo-sequence-based method to decompose the LF image into multiple views. Then all the views are efficiently represented using a 2-D coding structure. Around the 2-D coding structure, this paper has made the following key contributions to fully exploit the high correlations among various views.

1) Efficient encoding order. We first divide all the views in the 2-D coding structure into four quadrants, and encode the four quadrants one after another. In this way, except for the views in the border of two quadrants, each quadrant can be considered as independent of each other and thus the reference buffer size can be reduced as much as possible. Besides, inside each quadrant, the views are proposed to be encoded in a hierarchical coding structure in both horizontal and vertical directions to fully exploit the correlations among various views.
2) Distance based reference frame selection and motion vector (MV) scaling. For reference frame selection, we propose to use the distance between the current view and its reference views instead of the picture order count (POC) difference as the criterion for selecting better reference views for each inter view. The reference frames having the smaller distances with the current view will be put in the front of the reference frame lists. Besides, for MV scaling, the distance instead of the POC difference is also utilized as the scaling factor to improve the precision of the MV predictors.
3) Optimal bit allocation. We try to optimize the bit allocation for the proposed 2-D hierarchical coding structure through a two-pass coding scheme. In the first pass, we collect some statistics on the quality dependencies between different views, and derive the optimal λ ratios of various views in the proposed 2-D coding structure. The λ here is the Lagrange Multiplier, which determines the bitrate and distortion of each view. And in the second pass, the optimal λ ratios will be applied to the encoding process to optimize the coding efficiency.

Part of the work has already been published in [11]. In this paper, we give a more theoretical analysis of the proposed 2-D hierarchical coding framework, and we also provide an optimal bit allocation algorithm to further significantly improve the LF image coding efficiency. The experimental results show that the proposed algorithm can lead to three times of bitrate savings compared with the previous work.

This paper is organized as follows. Section II gives an overview of the related work. Section III will provide a detailed description of the proposed 2-D hierarchical coding structure with a limited number of frames. The distance based reference frame selection and MV scaling will also be described in

details in this section. Section IV will introduce the bit allocation algorithm for the proposed 2-D hierarchical coding structure in details. The experimental results on the proposed algorithms will be shown in Section V. Section VI concludes the whole paper.

II. Related Work

As mentioned in Section I, the LF image compression methods can be mainly divided into two categories. The first kind is the self-similarity based methods. The self-similarity here means that various views of the LF image are with very high correlations. The self-similarity based methods can be further divided into two groups. One group of them [12] proposed to use the so-called "image B-coder" by introducing the entire inter prediction scheme in HEVC into the context of intra prediction to make full use of the self-similarity. In the "image B-coder", the neighboring blocks belonging to same Coding Tree Unit (CTU) were set as list0 and the neighboring blocks belonging to the left CTU were set as list1. Both the uni-directional and bi-directional predictions are applied and the inter prediction directions are determined through the rate distortion optimization (RDO) for each block. Moreover, Li *et al.* [13] further extended this work to response for the 2016 ICME LF image compression challenge through investigating various rate-distortion ratios.

The other group of the self-similarity based methods tries to use the self-similarity compensated prediction and estimation to exploit the high correlations among various views. The self-similarity compensated prediction is very similar to the intra block copy (IBC) [9] in HEVC screen content coding [14]. The main difference is that the IBC uses the hash-based distortion metric in the motion estimation (ME) process while the sum of the absolute difference (SAD) is utilized in the ME process of the self- similarity compensated prediction. Conti *et al.* [15] first introduced the concept of self-similarity to H.264/AVC to compress the LF image in a more efficient way. This work was then extended to HEVC to take advantage of the flexible partitions [8]. In [8], the self similarity vectors were carefully predicted and encoded according to the characteristics of the LF image. Moreover, in [16], the bi-directional self-similarity compensated prediction and estimation were proposed to further improve the coding efficiency. In [17], the template matching was used to find multiple similar blocks to get a combined prediction block to improve the prediction accuracy.

The second kind of the LF image compression methods is the pseudo-sequence-based methods. The essence of the pseudo-sequence-based methods is decomposing the LF image into multiple views and compressing them like a sequence. The pseudo-sequence-based methods can also be further divided into two groups. The first group is the transform based methods. After decomposing the LF image into multiple views, 3D-Discrete Cosine Transform (3D-DCT) [18] [19] was applied to exploit the spatial redundancy as well as the temporal redundancy. Besides 3D-DCT, 3D-Discrete Wavelet Transform (3D-DWT) [20] [21] was also used to exploit the correlations among various views to better improve coding efficiency. Moreover, Elharar *et al.* [22] proposed a hybrid compression scheme which first applied a 2D DWT to each micro-image followed by a 2D DCT applied to sets of DWT coefficients from neighboring micro-images.

Along with the development of image and video coding standards, it has been demonstrated that the latest video coding standards with the block-based motion estimation and compensation are more efficient than the transform based methods. Therefore, the other group of the pseudo-sequence-based methods is based on the inter prediction of the newest video coding standard HEVC. For example, Perra and Assuncao [23] proposed to divide the LF image into multiple tiles and organize all these tiles sequentially into a sequence. Then the correlations among different frames were exploited using the HEVC inter prediction. Besides, Liu *et al.* [10] first proposed to decompose the LF image into multiple views according to the relative positions of the different micro-lenses. Then the multiple views are organized into a sequence according to their spatial position relationship. However, the coding orders, the reference frame management, and the bit allocation among various views are considered in a very coarse way thus the inter dependency among different views has not been fully exploited. Therefore, the rate-distortion (R-D) performance of the LF image compression scheme is far from the best.

Comparing the above mentioned two kinds of methods, Viola *et al.* [24] reported that the self-similarity based LF image compression methods cannot achieve the comparable performance with the pseudo-sequence-based methods due to their inflexibility to exploit the correlations among various views, especially in low bitrate case. Therefore, the proposed method in this paper also follows the approaches of the pseudo-sequence-based methods.

III. The Proposed 2-D Hierarchical Structure

The proposed 2-D hierarchical coding structure will be introduced in two aspects in the following two subsections. In Section III-A, we will introduce the basic concept of the proposed 2-D hierarchical coding structure within a limited number of reference frames. Then in Section III-B, we will describe the distance based reference frame selection and MV scaling algorithms under the 2-D hierarchical coding structure.

A. Basic Concept of the Proposed 2-D Hierarchical Structure

The LF image is first decomposed into multiple views following the approaches in [10]. Then some corner views, which are not beneficial for the overall LF image quality, are deleted from the original structure. After these steps, for the test LF images generated by Lytro as shown in [25], we can derive 165 views in total and organize all the views into a 2-D coding structure as shown in Fig. 2. It should be noted that the LF image generated by Lytro with the resolution of 7728×5368 is only used as an example to introduce the proposed 2-D hierarchical coding structures. The proposed algorithm can be easily extended to other kinds of LF images with various resolutions and a different number of views.

Then we try to find the view that has the highest correlation with the other views and code it as an intra frame. Here we define the correlation between the current view and all the other views

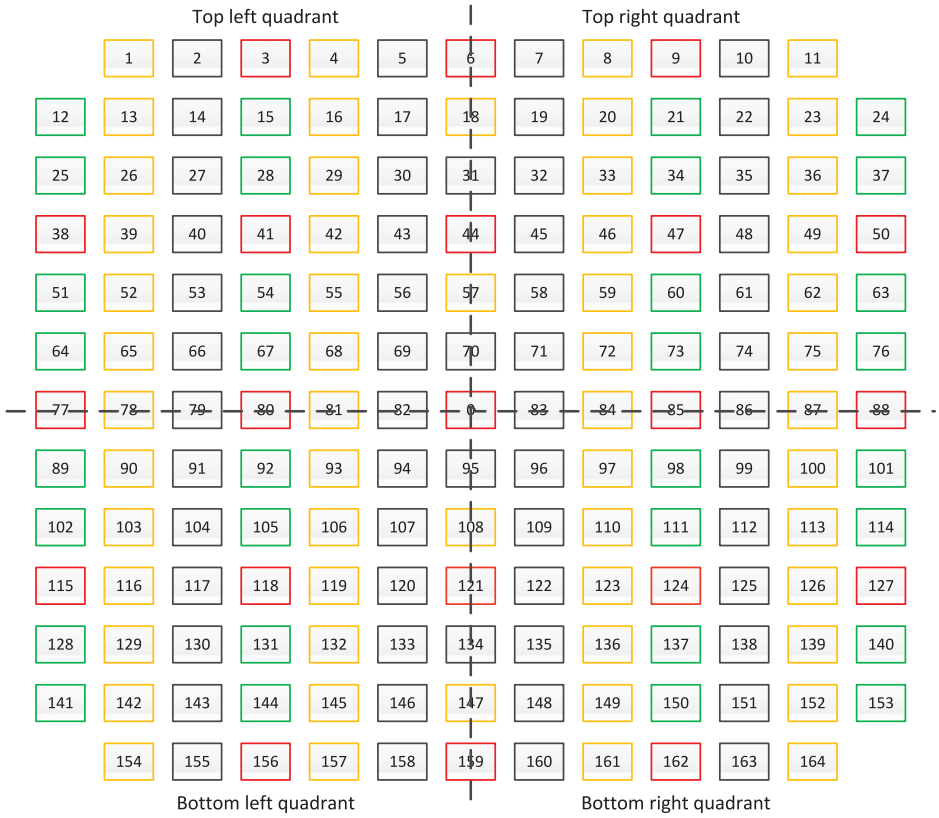


Fig. 2. This figure depicts the 13 × 13 views (excluding the 4 corner views) to be compressed. The views are assigned picture order count 0-164 shown in the figure. The views are divided into four quadrants.

TABLE I
THE TOP THREE SMALLEST AVERAGE DIFFERENCES BETWEEN THE CURRENT VIEW AND THE OTHER VIEWS FOR VARIOUS SEQUENCES

	Top 1		Top 2		Top 3	
Test images	POC	Diff	POC	Diff	POC	Diff
I01	70	9.27	0	9.27	83	9.28
I02	70	11.40	0	11.42	82	11.49
I03	0	10.92	70	10.95	83	10.95
I04	70	7.64	71	7.67	0	7.68
I05	0	8.92	69	8.92	82	8.93
I06	0	4.33	95	4.34	83	4.35
I07	83	6.29	0	6.31	71	6.33
I08	0	4.38	83	4.39	95	4.41
I09	0	12.73	83	12.75	95	12.77
I10	0	5.12	70	5.13	83	5.19
I11	0	9.64	70	9.67	83	9.69
I12	0	10.54	70	10.58	83	10.62

as the average frame difference per pixel between the current view and the other views. Table I shows the top three smallest average differences per pixel between the current view and the other views for various LF images. From the table, we can see that the average frame difference per pixel of the center view is the top one smallest for most test images and it is always within the top three smallest. Therefore, in our scheme, the center view is assigned with POC 0 and coded as an intra frame. The other views are assigned the POC sequentially from top left to bottom right as shown in Fig. 2 and coded as inter B frames. Note that the POC here is just a symbol to represent each view instead of the display order in usual videos.

After the POC is assigned to each view, then comes the determination of the coding orders of all the views and the reference relationships among various views. Here, we have not used the RDO theory to optimize the reference frame management. There are mainly two reasons for us to make such a choice. First, it has been shown in [30] that the RDO based reference frame management algorithm can only achieve negligibly better R-D performance compared with the 1-D hierarchical coding structure for the natural test sequences. Second, with the RDO based reference frame management, the encoder complexity will increase by several times under the 1-D hierarchical coding structure. Therefore, in this paper, we extend the 1-D hierarchical coding structure to construct a 2-D hierarchical coding structure to make full use of the inter correlations among different views instead of using the RDO based reference frame management. When designing the 2-D hierarchical coding structure, we try to simultaneously achieve a quite good coding efficiency and use a relatively small reference buffer.

As we know, in the 1-D hierarchical coding structure, the depth first coding order can lead to the least reference buffer size [26]. For example, as shown in Fig. 3, the encoding order of a typical 1-D hierarchical coding structure with group of pictures (GOP) size 16 is 0, 16, 8, 4, 2, 1, 3, 6, 5, 7, 12, 10, 9, 11, 14, 13, 15. And the minimum size of the reference buffer

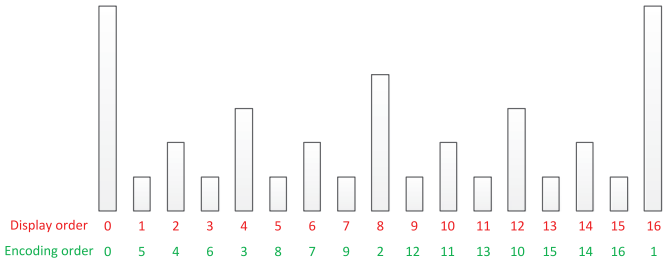


Fig. 3. The 1-D hierarchical coding structure with GOP size 16.

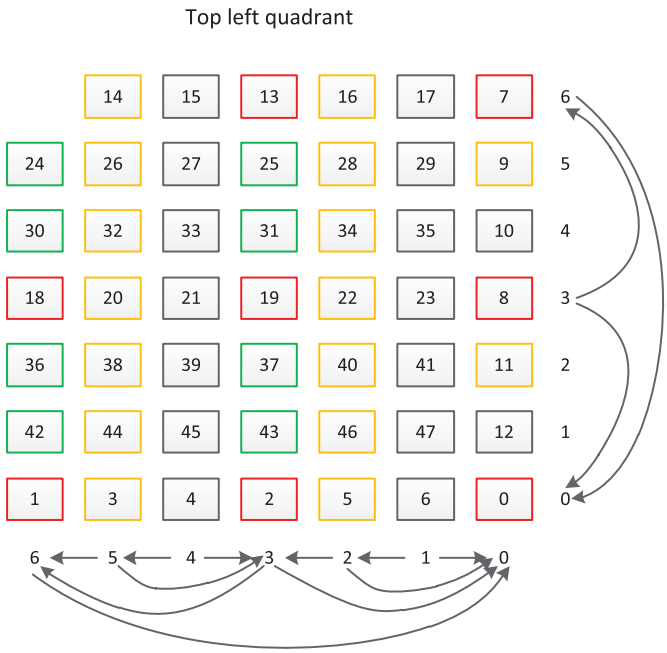


Fig. 4. The coding order of the top left quadrant.

is 5. The situation in the 2-D hierarchical coding structure is quite similar. To reduce the reference buffer size as much as possible, we first divide all the frames into four quadrants as shown in Fig. 2. Then each quadrant is coded in clockwise order from the top left quadrant, the top right quadrant, the bottom right quadrant, to the bottom left quadrant. In this way, except for the frames in the border of two quadrants, each quadrant can be considered as an independent one. Therefore, the reference frames belonging to only one quadrant can be pop out of the reference frame buffer as soon as possible to keep a relatively small reference buffer without influencing the coding efficiency.

Inside each quadrant, the depth first coding order will be used for both the horizontal and vertical directions to make full use of the correlations among various views. Take the top left quadrant as an example. The detailed encoding order can be seen from Fig. 4. In Fig. 4, the number inside each rectangle means the encoding order of each view. In both horizontal and vertical directions, we will follow the encoding order of 0, 6, 3, 5, 4, 2, 1. To be more specific, we will first encode the $0th$ row and the $0th$ column. Then according to the hierarchical coding structure in the vertical direction, the $6th$ row will be coded and followed by the $3rd$ row. Finally, we will encode the $5th$, $4th$, $2nd$, and $1st$ row. Inside each row or column, the order of 0, 6, 3, 5, 4, 2, 1 will also be used. It should also be noted that to guarantee the smallest reference buffer size, the coding order of the other three quadrants will be symmetrical to that of the top left quadrant. To be more specific, the encoding order for each row of the top right and bottom right quadrants will be from right to left, and the encoding order for various rows of the bottom right and bottom left quadrants will be from bottom to top. Take the row with frames 19, 20, 21, 22, 23, 24 as an example, the encoding order of the row will be from right to left as 24, 21, 23, 22, 20, 19. In this way, the frame 24 will only need to be stored in the reference buffer for the frame 23 and thus such a scheme can keep the reference frame buffer as small as possible.

After the encoding order is determined, then we will select the reference frames for each frame. As shown in Fig. 2, all the views are divided into four groups according to their frequencies to be referenced. The frequency means the times a reference frame can be referenced by other frames.

1) The frames with the red block. This kind of frames is the most frequently referenced frames. They are always stored in the reference buffer until the end of encoding of the current quadrant. All the frames including the to-be-encoded red block frames in the current quadrant can take the red block frames as reference. In the current quadrant, the existence of these frames can guarantee that all the frames have a good prediction.
2) The frames with the green block. This kind of frames is the second most frequently referenced frames. They will be referenced by the frames belonging to the current row in the same quadrant. For example, besides the red frames, the frame 26 can also take the frames 25 and 28 as references.
3) The frames with the yellow block. This kind of frames will only be referenced by the frame encoded immediately after them in the same quadrant. For example, the frame 27 can take the frame 26 as reference.
4) The frames with the black block. This kind of frames such as the frame 27 belongs to the non-reference frames.

It should be noted that since we are using a row-based coding order, to save the reference buffer size, the vertical references are much less than the horizontal references for most frames. For example, when encoding the frame 29, the immediate above frame 16 is unavailable under the proposed 2-D hierarchical coding structure.

Then we will analysis the reference buffer size needed to achieve such a reference frame structure. In the top left quadrant, according to our careful analysis, we find that the maximum number of reference frames appears when encoding the non-reference frames such as the frame 27. We should not only store the eight frequently used red reference frames in the reference buffer, but also the frames 26 and 28. Therefore, maximum 10 reference frames are needed for the top left quadrant. For the top right and bottom right quadrants, the situation is more complicated. As HEVC provides the constraint that the reference frames of the next frame can only be chosen from the reference frames of the current frame and the current frame itself [27], the frames 77 and 80 should always be stored in the reference buffer because they will be used as the reference frames for the

bottom left quadrant. Therefore, maximum 12 reference frames are needed for the top right and bottom right quadrants. The situation of the bottom left quadrant is the same as that of the top left quadrant, for which maximum 10 reference frames are needed. Therefore, in summary, the reference buffer size needed for encoding the entire pseudo sequence is 12.

B. Distance Based Reference Frame Selection and MV Scaling

According to the analysis in Section III-A, we know that the maximum number of reference frames is 12. If we apply all these reference frames for both list0 and list1, the encoder will traverse all the reference frames to obtain the best prediction block, which may increase the encoding complexity dramatically. Besides, the large reference index introduced by a large number of reference frames may also increase the overhead bits. Therefore, in this section, we first propose a distance based reference frame selection algorithm to reduce the overhead bits and decrease the encoding complexity. Then since the spatial positions of various views may also have influences on the MV scaling process in the merge [28] and advanced motion vector prediction modes, we also propose a spatial-coordinates-based MV scaling to further improve the coding efficiency.

In the 1-D hierarchical coding structure, the frames with the smaller POC differences are put in a relatively earlier position of the reference lists since they are nearer to the current frame and have larger possibilities to be referenced. However, in the proposed 2-D hierarchical coding structure, the POC is just a symbol to represent each view so the POC difference cannot reflect the distance between two frames. For example, as shown in Fig. 2, the POC difference between the frames 18 and 6 is 12, which is larger than the POC difference between the frames 18 and 15. However, the distance between the frames 18 and 6 is much smaller than that between the frames 18 and 15. Therefore, we need to first obtain the distances among various views before selecting the suitable reference frames instead of using the POC differences. In our implementation, we establish a coordinate system to derive the spatial coordinates of all the views and then use the spatial coordinates to calculate the distances among various views. The spatial coordinate of the most top left position of Fig. 2 is set as $(0, 0)$, and the right and down directions are set as positive. For example, the spatial coordinates of the frame 0 and the frame 1 is $(6, 6)$ and $(1, 0)$, respectively. Then we can derive the correspondence between the POC and the spatial coordinate of each frame (x, y) as follows.

$$x = \begin{cases} 6 & \text{if POC} = 0 \\ \text{POC}\%13 & \text{else if POC} \leq 11 \\ (\text{POC}+1)\%13 & \text{else if POC} \leq 82 \\ (\text{POC}+2)\%13 & \text{else if POC} \leq 153 \\ (\text{POC}+3)\%13 & \text{otherwise} \end{cases} \quad (1)$$

$$y = \begin{cases} 6 & \text{if POC} = 0 \\ \text{POC}/13 & \text{else if POC} \leq 11 \\ (\text{POC}+1)/13 & \text{else if POC} \leq 82 \\ (\text{POC}+2)/13 & \text{else if POC} \leq 153 \\ (\text{POC}+3)/13 & \text{otherwise} \end{cases} \quad (2)$$

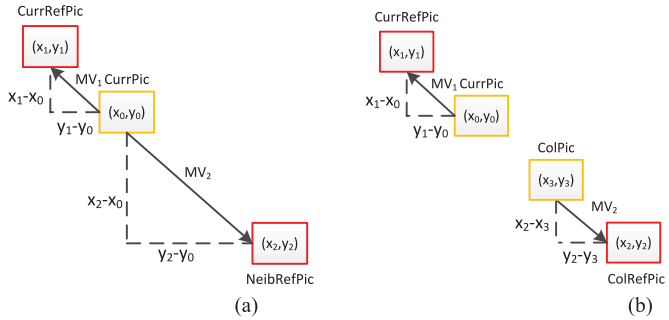


Fig. 5. Distance based MV scaling. (a) Spatial case. (b) Temporal case.

After the spatial coordinates are determined, we can easily calculate the distance between the frames (x_1, y_1) and (x_2, y_2) through the euclidean distance.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (3)$$

Then we will construct both the list0 and list1 according to the distances between the current frame and its reference frames. In the 1-D hierarchical coding structure, the forward (smaller POC compared with the current frame) and backward (larger POC compared with the current frame) reference frames are put into list0 and list1, respectively. Similarly, under the 2-D hierarchical coding structure, we should first define the forward and backward directions. In our implementation, the above frames are all treated as the forward frames, and the below frames are all treated as the backward frames. For example, for the frame 17, the frames 1 to 16 are considered as the forward frames, and all the other frames including the frame 0 are considered as the backward frames. According to the above introduction, the totally available reference frames for the frame 17 are frames 16, 15, 6, 3, 38, 41, 44, 77, 80, and 0. If we set the number of reference frames in both lists as 4, according to the distances between the current frame and its reference frames in increasing order, the reference frames in list0 will be 16, 6, 15, and 3, and the reference frames in list1 will be 44, 41, 0, and 80.

Besides the reference frame selection, the spatial coordinates may also have significant influences on the MV scaling operations in both merge and advanced motion vector prediction modes. The MV scaling operations are performed when the spatial neighboring blocks or temporal co-located blocks are pointing to a different reference frame from the current block. The MV scaling can be divided into two kinds: the spatial and temporal MV scaling. In 2-D hierarchical coding structure, we should perform MV scaling based on the distance in x and y directions separately instead of POC.

The detailed processes are presented in Fig. 5. For the spatial case, the spatial coordinate of the current frame is (x_0, y_0), the spatial coordinate of the current reference frame is (x_1, y_1), and the spatial coordinate of the reference frame of the neighboring block is (x_2, y_2). The MV of the current block is $(\text{MV}_{1,x}, \text{MV}_{1,y})$, and the MV of the neighboring block is $(\text{MV}_{2,x}, \text{MV}_{2,y})$. Assuming that the motions among various frames are uniform, the spatial MV scaling process can be

derived as follows.

$$\mathrm{MV}_{1,x} = \frac{\mathrm{MV}_{2,x}}{x_2 - x_0} \times (x_1 - x_0) \tag{4}$$

$$\mathrm{MV}_{1,y} = \frac{\mathrm{MV}_{2,y}}{y_2 - y_0} \times (y_1 - y_0) \tag{5}$$

It should be noted that the values of $x_2 - x_0$ and $y_2 - y_0$ are zero when the current frame and the neighboring reference frame are in the same row or column. In this case, the spatial scaling operations are not applied.

For the temporal case, except for the current frame and its reference frame, there are temporal co-located frame and its corresponding reference frame whose spatial coordinates are (x_3, y_3) and (x_2, y_2), respectively. The MV of the current block is $(\mathrm{MV}_{1,x}, \mathrm{MV}_{1,y})$, and the MV of the co-located block is $(\mathrm{MV}_{2,x}, \mathrm{MV}_{2,y})$. Assuming that the motions among various frames are uniform, the temporal MV scaling process can be derived as follows.

$$\mathrm{MV}_{1,x} = \frac{\mathrm{MV}_{2,x}}{x_2 - x_3} \times (x_1 - x_0) \tag{6}$$

$$\mathrm{MV}_{1,y} = \frac{\mathrm{MV}_{2,y}}{y_2 - y_3} \times (y_1 - y_0) \tag{7}$$

Similar to the spatial MV scaling process, the values of $x_2 - x_3$ and $y_2 - y_3$ are zero when the co-located frame and the co-located reference frame are in the same row or column. In this case, the temporal scaling operations are not applied.

IV. The Proposed Optimal Bit Allocation Schemes

In the proposed 2-D hierarchical coding structure, the LF image is decomposed into multiple views, and the multiple views are organized into a pseudo sequence to better exploit the correlations among multiple views. Therefore, minimizing the distortion of the LF image under the bitrate constraint becomes equivalent to minimizing the distortion of the decomposed pseudo sequence with the same bitrate constraint. As mentioned in the Section III-A, all the frames are divided into four groups according to their frequencies to be referenced. In Fig. 2, the frames with the red rectangle are referenced the most while the frames with the black rectangle are non-reference frames. The quality of the frames which are referenced the most has the largest influence on the quality of the whole sequence and should be assigned the largest number of bits, while the quality of the non-reference frames has the smallest influence on the quality of the whole sequence and should be assigned the least number of bits. In this section, we will investigate the influence of various frames on the whole sequence, and then propose an optimal bit allocation algorithm to assign each frame with the suitable number of bits.

In the recent researches, Li *et al.* [29], [30] proposed that the Lagrange Multiplier λ instead of the quantization parameter (QP) was the key factor to determine the bits and distortion for each frame. In other words, to assign a suitable number of bits to each frame is equivalent to determining a suitable λ_i for each frame. Therefore, the optimization target of the proposed bit allocation algorithm is to determine a suitable λ_i for each frame to optimize the overall quality of all the frames under the target bits constraint,

$$\min_{\lambda_1,\lambda_2,\ldots,\lambda_N} \sum_{i=1}^{N} D_i, s.t. \sum_{i=1}^{N} R_i \le R_t \tag{8}$$

where the λ_i is the λ for the frame i, which determines the optimization target of the frame i. The D_i and R_i are the distortion and bits for the frame i, respectively. N is the total number of frames in the pseudo sequence. Note that, different from the traditional approach [31] to optimize the bit allocation, the proposed method enrolls λ_i for each picture as the optimization parameter. The constrained problem can be converted to the following unconstrained problem,

$$\min_{\lambda_1,\lambda_2,\ldots,\lambda_N} \sum_{i=1}^{N} D_i + \lambda \sum_{i=1}^{N} R_i \tag{9}$$

where the λ is the Lagrange Multiplier of the optimization problem, and determines the optimization target of the whole pseudo sequence. In the following, we will demonstrate that the λ is equal to the λ_i of the non-reference frames.

To solve (9), the common used Lagrange method is applied. Then we can derive the following equation by setting the derivative of the total cost to 0,

$$\frac{\partial \sum_{i=1}^{N} D_i}{\partial \lambda_j} + \lambda \frac{\partial \sum_{i=1}^{N} R_i}{\partial \lambda_j} = 0, \quad j = 1, 2, \ldots N \tag{10}$$

Since the λ_j is related to the distortion D_j, (10) can be converted to the following formula,

$$\frac{\partial \sum_{i=1}^{N} D_i}{\partial D_j} \cdot \frac{\partial D_j}{\partial \lambda_j} + \lambda \frac{\partial \sum_{i=1}^{N} R_i}{\partial D_j} \cdot \frac{\partial D_j}{\partial \lambda_j} = 0, \quad j = 1, 2, \ldots N \tag{11}$$

As the quality of the current frame will only have influence on itself and its subsequent frames, (11) can be rewritten in another form by ignoring the term $\frac{\partial D_j}{\partial \lambda_j}$,

$$\frac{\partial \sum_{i=j}^{N} D_i}{\partial D_j} + \lambda \frac{\partial \sum_{i=j}^{N} R_i}{\partial D_j} = 0, \quad j = 1, 2, \ldots N \tag{12}$$

Then both $\frac{\partial \sum_{i=j}^{N} D_i}{\partial D_j}$ and $\frac{\partial \sum_{i=j}^{N} R_i}{\partial D_j}$ can be written as the sum of the influence to itself and its subsequent frames,

$$\frac{\partial \sum_{i=j}^{N} D_i}{\partial D_j} = 1 + \frac{\partial \sum_{i=j+1}^{N} D_i}{\partial D_j} \tag{13}$$

$$\frac{\partial \sum_{i=j}^{N} R_i}{\partial D_j} = \frac{\partial R_j}{\partial D_j} + \frac{\partial \sum_{i=j+1}^{N} R_i}{\partial D_j} \tag{14}$$

Since the λ_j is the slope of the R-D curve for each frame, we can obtain the following formula,

$$\frac{\partial R_j}{\partial D_j} = \frac{1}{\frac{\partial D_j}{\partial R_j}} = -\frac{1}{\lambda_j} \tag{15}$$

By substituting (15), (13), and (14) into (12), we can obtain the equation as follows,

$$\left(1+\frac{\partial\sum_{i=j+1}^{N}D_i}{\partial D_j}\right)+\lambda\left(-\frac{1}{\lambda_j}+\frac{\partial\sum_{i=j+1}^{N}R_i}{\partial D_j}\right)=0 \tag{16}$$

Here we define the R-D cost of each frame J_i as

$$J_i \triangleq D_i+\lambda R_i, \quad i=1,2,...N \tag{17}$$

It should be emphasized that for each frame, to calculate the R-D cost, the Lagrange Multiplier used is the sequence level λ instead of the λ_i for each specified frame. Then we can solve (16) as

$$\lambda_j=\frac{\lambda}{(1+\frac{\partial\sum_{i=j+1}^{N}J_i}{\partial D_j})}\triangleq\frac{\lambda}{1+\omega_j} \tag{18}$$

To make the physical interpretation of (18) more obvious, we rewrite (18) as

$$\frac{\lambda_i}{\lambda_j}=\frac{1+\omega_j}{1+\omega_i} \tag{19}$$

Equation (19) means that the λ of different frames should be inversely proportional to its influence to the R-D cost of the whole sequence to achieve the optimal R-D performance. The larger the influence of the current frame on the whole sequence, the smaller the λ of the current frame will be. The small λ will result in a small distortion and thus a large number of bits assigned to the current frame. As we all know, a large number of bits assigned to the frame having a far-reaching influence on the whole sequence is beneficial the overall performance.

Then the only problem becomes how to determine the term w_j in (18). We know that the non-reference frames have no influence to the subsequent frames since they will not be referenced by any frames. Therefore, for the non-reference frames, the value of w_j will be equal to 0. From (18), we can easily see that sequence level λ is equal to the λ_j of the non-reference frames [32]. This work explicitly demonstrates that the sequence level λ is equal to the λ_j of the non-reference frames.

For the reference frames, we use a two-pass scheme similar to the methods in [31] to optimize the bit allocation problem. In the first encoding pass, we use the simple λ ratio between various frames in [10] as the initial bits ratio to generate the anchor R-D cost J_{j0} and the distortion D_{j0}. Then we change the QP and λ of the frame j to generate the test R-D cost J_{j1} and distortion D_{j1}. Both the J_{j0} and J_{j1} only include the R-D cost of the frames coded after the frame j since the frames coded before the frame j will not change. It should be noted that the λ used to calculate the R-D cost is the sequence level λ according to (17). Besides, the QP of the frame j is decreased by 5 to guarantee that the frame j will be with an obvious distortion change ΔD_j [31]. The number 5 is chosen according to our experience. It is neither too small to make a difference on the R-D cost of the entire pseudo sequence, nor too large to be not within a normal range. In this way, we can easily obtain the value of delta R-D cost ΔJ_j after changing the distortion of the test frame with ΔD_j. Then we can obtain an approximate value for ω_j.

$$\omega_j\approx\frac{\Delta J_j}{\Delta D_j}=\frac{J_{j1}-J_{j0}}{D_{j0}-D_{j0}} \tag{20}$$

In the second encoding pass, we can use the ω_j obtained in the first pass to calculate the λ_j of each frame according to (18). The QP is then determined through the following equation according to experience [33],

$$QP=4.3281\times\log(\lambda)+14.4329 \tag{21}$$

After the determination of the λ and QP for each frame, we can finish the entire encoding process through RDO.

V. Experimental Results

A. Simulation Setup

The proposed pseudo sequence based 2-D hierarchical coding structure for LF image compression is implemented in HM-16.7 [34]. The previous pseudo-sequence-based LF image compression method proposed in [10] is used as the anchor to demonstrate the effectiveness of the proposed algorithm. Since the bitrates generated by various LF image compression algorithms are different, the Bjontegaard-Delta rate (BD-rate) [35] is used to measure the performance for a fair comparison. We use both the Y-PSNR and YUV-PSNR between the original LF image and the reconstructed LF image shown in [4] as an objective quality metric. The test images used are also specified in [4].

Since an optimal bit allocation algorithm is proposed for the 2-D hierarchical coding structure, we use the following method to guarantee that the optimization targets of the proposed algorithm and the anchor are the same. We first generate the anchor using the LF image compression algorithm in [10] with intra frame QPs 15, 20, 25, and 30. Then the λ of the non-reference frames is recorded and used as the λ of the non-reference frames for the proposed algorithm. The λs of the reference frames are then calculated using (18).

The performance of the proposed algorithm will be introduced in the following subsections. We will first present some experimental results on the overall framework. Then we will analyze the benefits from the following three aspects: the proposed distance based reference frame selection, spatial-coordinates-based MV scaling, and the optimal bit allocation algorithms.

B. The Performance of the Overall Framework

The performance of the overall framework is shown in Table II. From Table II, we can see that the proposed algorithm can provide on average 18.3% and 19.0% bitrate savings in Y and YUV components, respectively. For the sequence I06-rawLF, the R-D performance improvements can be as high as 28.4% and 28.7% accordingly. It should also be mentioned that the bitrate savings brought by the proposed algorithm are quite consistent. We can achieve significant bitrate savings for all the test sequences without even one exception. The experimental results obviously demonstrate that the proposed LF image compression framework can well exploit the correlations among various views, and thus can significantly improve the

TABLE II
THE PERFORMANCE OF THE OVERALL FRAMEWORK

test images	Y-BDrate	YUV-BDrate	Enc. Time	Dec. Time
I01	–20.5%	–21.7%	94.3%	102.5%
I02	–16.4%	–18.2%	93.8%	98.1%
I03	–22.7%	–23.8%	93.9%	99.1%
I04	–16.6%	–16.0%	94.7%	100.1%
I05	–19.6%	–21.1%	91.7%	98.0%
I06	–28.4%	–28.7%	92.3%	100.6%
I07	–17.0%	–17.9%	92.5%	97.8%
I08	–12.9%	–12.7%	92.2%	100.6%
I09	–22.0%	–23.1%	94.1%	99.2%
I10	–15.1%	–17.1%	96.1%	103.8%
I11	–6.5%	–6.1%	93.1%	100.1%
I12	–8.0%	–8.7%	93.7%	102.7%
Average	–18.3%	–19.0%	94.0%	101.5%

TABLE III
THE PERFORMANCE OF THE DISTANCE BASED REFERENCE FRAME SELECTION

test images	Y-BDrate	YUV-BDrate	Enc. Time	Dec. Time
I01	–6.4%	–6.5%	94.2%	99.1%
I02	–4.6%	–4.6%	94.6%	97.1%
I03	–4.7%	–4.7%	93.7%	97.2%
I04	–2.8%	–2.1%	93.8%	95.6%
I05	–5.6%	–5.6%	93.4%	96.9%
I06	–13.8%	–13.7%	92.5%	97.5%
I07	–8.2%	–8.3%	93.7%	96.7%
I08	–12.2%	–12.1%	93.6%	99.5%
I09	–1.7%	–1.7%	95.1%	100.2%
I10	–9.2%	–9.1%	94.7%	100.1%
I11	–1.1%	–1.0%	95.4%	101.5%
I12	0.9%	0.7%	95.8%	99.7%
Average	–5.8%	–5.7%	94.3%	99.2%

R-D performances for both the Y and YUV components. Besides, note that the performance improvement provided by the overall framework consists of three parts: the distance based reference frame management, the distance based MV scaling, and the optimal bit allocation algorithm. The performance improvement shown in Table II is not provided individually by the proposed 2-D hierarchical coding structure since the optimal bit allocation algorithm is not enabled in the anchor. We will introduce the performances of proposed 2-D hierarchical structure and the optimal bit allocation individually in the next section.

For the encoding complexity, since the proposed 2-D hierarchical coding structure uses less reference frames compared with the anchor for some frames, the proposed algorithm presents less encoding time compared with the anchor. Note that the complexity of the first coding pass to calculate the optimal bit allocation ratios among various views is excluded since we mainly try to show the encoding time of the proposed 2-D hierarchical coding structure. The complexity of the first coding pass can be quite high since we need to encode the sequence multiple times to obtain the optimal bit allocation ratios. For the decoding complexity, since almost no changes are applied to the decoder, the proposed algorithm shows similar decoding time compared with the anchor.

C. The Performance and Analysis of the Separate Three Parts

1) The Proposed Distance Based Reference Frame Selection: The performance of the proposed distance based reference frame selection combined with the 2-D hierarchical coding structure is illustrated in Table III. From Table III, we can see that an average of 5.8% and 5.7% bitrate savings can be achieved through the proposed distance based reference frame selection. The proposed algorithm can achieve about $1/3$ of the total performance improvement. The performance improvement mainly comes from the fact that the distance based reference frame selection method can always guarantee a reference frame with relatively small distance with the current frame to be used as a real reference frame. Besides, compared with the previous pseudo-sequence-based LF image compression method, the proposed reference frame management algorithm can not only improve the R-D performance, but also save the reference frames to be stored in the reference frame buffer. The size of the reference buffer is only 12 through the effective management of the proposed algorithm. However, it can be as large as 51 under the algorithm in [10]. In summary, the proposed algorithm can achieve a better use of the reference frame buffer within a limited number of reference frames.

TABLE IV
THE PERFORMANCE OF THE DISTANCE BASED MV SCALING

test images	Y-BDrate	YUV-BDrate	Enc. Time	Dec. Time
I01	–0.5%	–0.4%	102.1%	100.2%
I02	–0.6%	–0.6%	102.0%	98.9%
I03	–0.6%	–0.5%	100.6%	98.7%
I04	0.0%	0.0%	100.9%	97.5%
I05	–0.8%	–0.7%	100.7%	97.8%
I06	–0.5%	–0.5%	100.0%	94.7%
I07	–0.8%	–0.8%	100.7%	98.1%
I08	–0.1%	0.0%	102.1%	98.6%
I09	–0.3%	–0.3%	102.0%	99.5%
I10	–0.1%	–0.2%	101.0%	97.9%
I11	–1.9%	–1.9%	101.8%	99.7%
I12	–0.9%	–0.9%	101.8%	100.0%
Average	–0.6%	–0.6%	101.2%	98.3%

2) The Proposed Distance Based MV Scaling: The performance of the proposed distance based MV scaling individually is illustrated in Table IV. From Table IV, we can see that an average of 0.6% and 0.6% bitrate savings on Y and YUV components can be achieved through the proposed distance based MV scaling. Since a LF image is with very dense views, most MVs between different views are very small. That is the reason why the performance improvement brought by the proposed distance based MV scaling method is limited. The performance of the combined distance-based reference selection and MV scaling algorithms are illustrated in Table V. From Table V, we can see that an average of 6.5% and 6.5% R-D performance improvement can be obtained by combining these two algorithms. The experimental results obviously demonstrate that the combination of these two algorithms can lead to even better R-D performance.

3) The Proposed Optimal Bit Allocation Algorithm: The performance of the proposed optimal bit allocation algorithm

TABLE V
THE PERFORMANCE OF COMBINED DISTANCE BASED REFERENCE FRAME SELECTION AND MV SCALING

test images	Y-BDrate	YUV-BDrate	Enc. Time	Dec. Time
I01	–7.1%	–7.1%	95.4%	98.4%
I02	–5.4%	–6.1%	95.1%	98.1%
I03	–5.4%	–5.5%	94.8%	99.9%
I04	–3.4%	–2.6%	94.9%	97.7%
I05	–6.5%	–6.4%	94.4%	99.0%
I06	–14.2%	–14.2%	94.0%	98.6%
I07	–8.6%	–8.7%	95.2%	99.5%
I08	–12.9%	–12.7%	94.3%	98.4%
I09	–2.6%	–2.6%	96.0%	100.1%
I10	–9.8%	–9.7%	96.2%	98.7%
I11	–2.0%	–1.9%	96.5%	101.6%
I12	–0.4%	–0.6%	96.6%	99.7%
Average	–6.5%	–6.5%	95.4%	99.8%

TABLE VI
THE PERFORMANCE OF THE OPTIMAL BIT ALLOCATION ALGORITHM

test images	Y-BDrate	YUV-BDrate	Enc. Time	Dec. Time
I01	–13.9%	–15.2%	106.7%	108.8%
I02	–10.9%	–12.4%	106.1%	106.1%
I03	–17.7%	–18.6%	104.5%	106.5%
I04	–12.9%	–12.8%	105.3%	107.4%
I05	–13.7%	–15.1%	101.7%	101.5%
I06	–19.9%	–20.1%	101.5%	98.3%
I07	–10.2%	–10.7%	101.9%	102.9%
I08	–13.1%	–13.7%	101.0%	100.0%
I09	–12.5%	–14.2%	103.5%	102.5%
I10	–18.9%	–18.8%	105.9%	108.3%
I11	–8.6%	–8.9%	100.7%	99.6%
I12	–6.6%	–7.1%	102.6%	103.1%
Average	–13.2%	–14.0%	103.6%	104.1%

on the coding structure of the anchor is shown in Table VI. From Table VI, we can see that the proposed bit allocation algorithm can achieve an average of 13.2% and 14.0% on the coding structure of the anchor. Also, through the comparison of Tables V and II, we can see that the proposed bit allocation algorithm can achieve also over 10% coding gains under the proposed 2-D hierarchical coding structure. The experimental results obviously demonstrate that the optimal bit allocation algorithm can achieve very significant bitrate savings under both coding structures.

Some typical examples of the QPs of various views are shown in Fig. 6. First, we can see from Fig. 6 that the QP of the first encoded view is much lower than the other views. Under the proposed 2-D reference structure, the first encoded view can be referenced by all the views in all four quadrants. Therefore, the coding quality of the first view has a far-reaching influence on the whole test sequence. That is why the first view should be assigned the largest number of bits and coded using a much smaller QP compared with the other views. Second, we can obviously see from Fig. 6 that the optimal bit allocation algorithm can assign reasonable QPs to various views according to the characteristics of various LF images. The characteristic of the LF image I11 is complex and the correlations between among various views are less compared with the LF image I06.

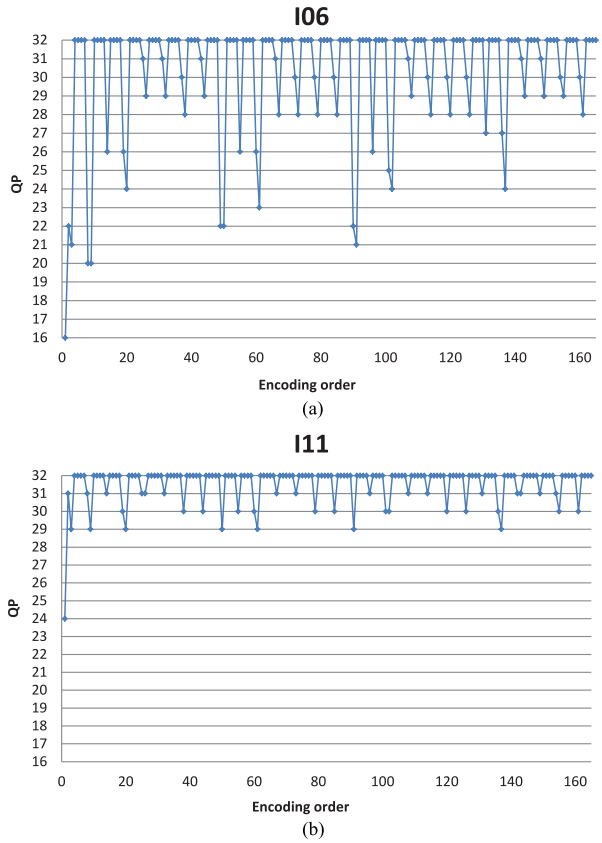


Fig. 6. The QPs of various views after bit allocation. (a) 106. (b) I11.

Therefore, the differences of QPs of various views in the LF image I11 are much smaller than that in the LF image I06. The LF image characteristic adaptation feature of the proposed bit allocation algorithm is the reason why it can bring so significant R-D performance improvement.

D. Some Examples of the R-D Curves

Some examples of the R-D curves are shown in Fig. 7. From these R-D curves, we can see that the proposed distance based reference frame selection and MV scaling can bring obvious R-D performance improvements compared with the anchor. It can also be seen from these R-D curves that the distance based reference frame selection and MV scaling can mainly reduce the bitrate instead of increasing the PSNR and the performance improvements mainly come from the low bitrate case. We can also see from these figures that the proposed optimal bit allocation algorithm can improve the R-D performance significantly no matter in the low bitrate or high bitrate cases. The proposed optimal bit allocation algorithm can bring consistent bitrate savings in both low bitrate and high bitrate cases.

E. The Comparison With the JPEG Under the Fixed Compression Ratios

Since the bitrates shown in Fig. 7 vary significantly from one image to another, we have also adjusted the QP for each test image to achieve compression ratios 10:1, 20:1, 50:1, and 100:1

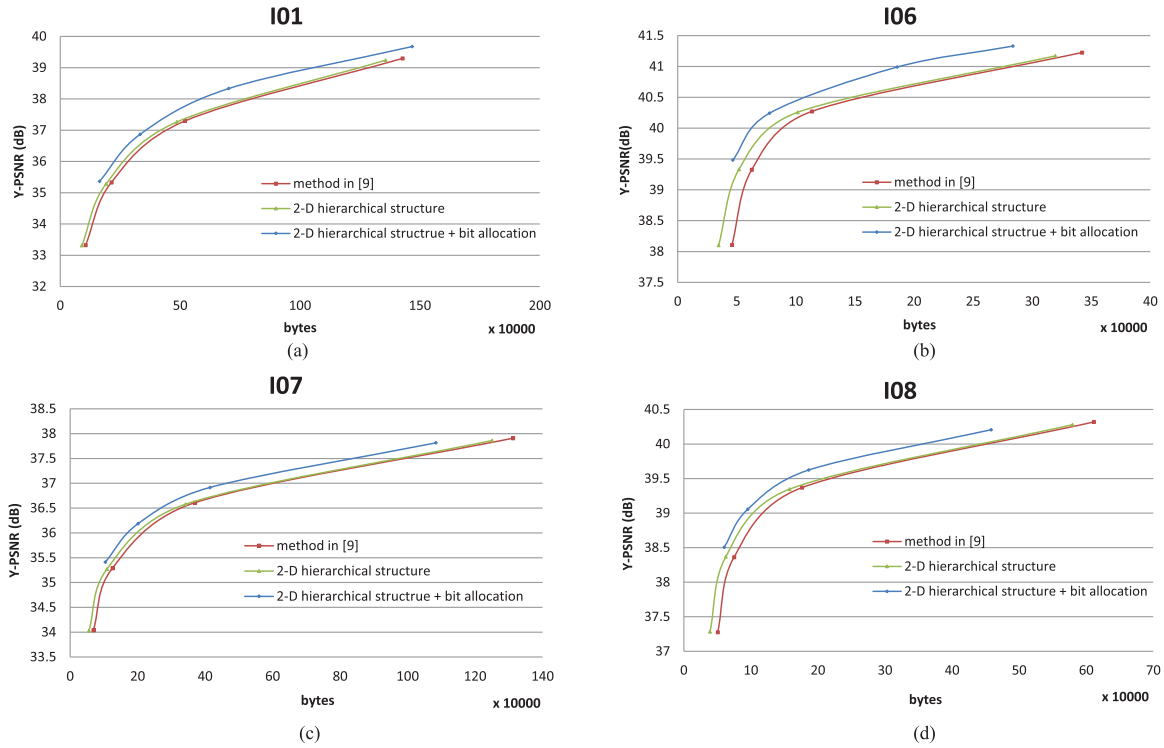


Fig. 7. Some examples of the R-D curves. (a) I01 (b) 106. (c) 107. (d) 108.

TABLE VII
THE BASE QP SETTING AND BD-PSNR OF THE PROPOSED ALGORITHM COMPARED WITH JPEG

test images	Base QPs	BD-PSNR (dB)
I01	9, 12, 16, 22	5.091
I02	10, 14, 19, 24	5.524
I03	10, 14, 19, 24	4.658
I04	8, 12, 16, 21	4.305
I05	10, 13, 16, 20	4.542
I06	5, 8, 10, 13	4.039
I07	7, 11, 14, 19	3.024
I08	5, 8, 10, 14	3.722
I09	9, 13, 16, 22	6.581
I10	7, 10, 14, 20	3.473
I11	10, 14, 17, 19	5.467
I12	8, 12, 16, 21	5.262
	Average	4.641

to compare with JPEG for a more convenient comparison for the future work. Table VII summarizes the QP setting and the BD-PSNR of the proposed algorithm compared with JPEG. From Table VII, we can see that the proposed 2-D hierarchical coding structure combined with the optimal bit allocation algorithm can achieve an average of 4.641dB BD-PSNR improvement compared with JPEG. Also, according to our observation, the BD-PSNR improvement is much more significant in low bitrate case compared with high bitrate case.

F. Summary of the Experimental Results

1) The proposed 2-D hierarchical coding structure, as well as the optimal bit allocation algorithm, can well exploit the correlations among various views and bring over 18% bitrate savings compared with the previous pseudo-sequence-based LF image compression method.
2) The proposed distance based reference frame selection combined with the 2-D hierarchical coding structure can bring about 6% R-D performance improvements.
3) The proposed distance based MV scaling method can bring about 0.6% bitrate savings.
4) The proposed optimal bit allocation algorithm for the 2-D hierarchical coding structure can bring over 10% R-D performance improvements for the LF-image compression due to its LF image content adaptive feature.

VI. CONCLUSION

In this paper, we propose a novel pseudo sequence based 2-D hierarchical coding structure for light-field image compression. The proposed 2-D hierarchical coding structure mainly has three key contributions. First, we propose a 2-D hierarchical coding structure with a limited number of reference frames. To be more specific, we divide all the views into four quadrants, and all the views are encoded one quadrant after another to reduce the reference buffer size as much as possible. Inside each quadrant, all the views are encoded hierarchically in both horizontal and vertical directions to fully exploit the correlations between different views. Second, we propose to use the distance between the current view and its reference views instead of the picture order count as the criterion for selecting better reference frames for each inter view. The distance between different views is also applied to the motion vector scaling process to obtain more accurate motion vector predictors. Third, an optimal bit

allocation algorithm for the proposed 2-D hierarchical structure is proposed to further exploit the inter correlations among various views and improve coding efficiency. The whole scheme is implemented in the reference software of High Efficiency Video Coding. The experimental results demonstrate that the proposed novel pseudo-sequence-based 2-D hierarchical structure can achieve on average 18.3% bit-rate savings compared with the previous pseudo-sequence-based LF image compression method.

In this work, both the coding order and the reference structure are determined according to our experience. Only the bit allocation is optimized using the rate distortion optimization theory. In the future, we will try to optimize the combinations of the coding order, the reference structure, and the bit allocation, using the fundamental rate distortion optimization theory. Besides, since two-pass encoding is needed to obtain the specific bit allocation for each sequence, we will try to follow the approaches in [36] and [37] to derive a one-pass method to achieve an optimal bit allocation in the future.

REFERENCES

[1] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Annu. Conf. Comput. Graph. Interactive Tech.* (SIGGRAPH '96). New York, NY, USA: ACM, 1996, pp. 31–42.

[2] L. Xu *et al.*, "FlyCap: Markerless motion capture using multiple autonomous flying cameras," *IEEE Trans. Vis. Comput. Graphics*, 2017.

[3] Light Field Tool box 0.4. 2015. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/49683-light-field-toolbox-v0-4

[4] M. Rerabek, T. Bruylants, T. Ebrahimi, F. Pereira, and P. Schelkens, "CFP of the ICME light field compression challenge," in *Proc. 2016 IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2016.

[5] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Trans. Consum. Electron.*, vol. 38, no. 1, pp. xviii–xxxiv, Feb. 1992.

[6] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[7] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[8] C. Conti, L. D. Soares, and P. Nunes, "HEVC-based 3D holoscopic video coding using self-similarity compensated prediction," *Signal Process.: Image Commun.*, vol. 42, pp. 59–78, Mar. 2016.

[9] X. Xu *et al.*, "Intra block copy in HEVC screen content coding extensions," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 6, no. 4, pp. 409–419, Dec. 2016.

[10] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *Proc. 2016 IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2016, pp. 1–4.

[11] L. Li, Z. Li, B. Li, D. Liu, and H. Li, "Pseudo sequence based 2-D hierarchical coding structure for light-field image compression," *2017 Data Compression Conf. (DCC)*, Snowbird, UT, 2017, pp. 131–140.

[12] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Coding of focused plenoptic contents by displacement intra prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 7, pp. 1308–1319, Jul. 2016.

[13] Y. Li, R. Olsson, and M. Sjöström, "Compression of unfocused plenoptic images using a displacement intra prediction," in *Proc. 2016 IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2016, pp. 1–4.

[14] J. Xu, R. Joshi, and R. A. Cohen, "Overview of the emerging HEVC screen content coding extension," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 50–62, Jan. 2016.

[15] C. Conti, J. Lino, P. Nunes, L. D. Soares, and P. L. Correia, "Spatial prediction based on self-similarity compensation for 3D holoscopic image and video coding," in *Proc. 2011 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 961–964.

[16] C. Conti, P. Nunes, and L. D. Soares, "HEVC-based light field image coding with bi-predicted self-similarity compensation," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2016, pp. 1–4.

[17] R. Monteiro *et al.*, "Light field HEVC-based image coding using locally linear embedding and self-similarity compensated prediction," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2016, pp. 1–4.

[18] R. Zaharia, A. Aggoun, and M. McCormick, "Adaptive 3D-DCT compression algorithm for continuous parallax 3D integral imaging," *Signal Process.: Image Commun.*, vol. 17, no. 3, pp. 231–242, 2002.

[19] A. Aggoun, "A 3D DCT compression algorithm for omnidirectional integral images," in *Proc. 2006 IEEE Int. Conf. Acoust., Speech, Signal Process. Proc.*, vol. 2, May 2006, pp. 517–520.

[20] A. Aggoun and M. Mazri, "Wavelet-based compression algorithm for still omnidirectional 3D integral images," *Signal, Image Video Process.*, vol. 2, no. 2, pp. 141–153, 2008.

[21] A. Aggoun, "Compression of 3D integral images using 3D wavelet transform," *J. Display Technol.*, vol. 7, no. 11, pp. 586–592, Nov. 2011.

[22] E. Elharar, A. Stern, O. Hadar, and B. Javidi, "A hybrid compression method for integral images using discrete wavelet transform and discrete cosine transform," *J. Display Technol.*, vol. 3, no. 3, pp. 321–325, Sep. 2007.

[23] C. Perra and P. Assuncao, "High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement," in *Proc. 2016 IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2016, pp. 1–4.

[24] I. Viola, M. Rerabek, T. Bruylants, P. Schelkens, F. Pereira, and T. Ebrahimi, "Objective and subjective evaluation of light field image compression algorithms," in *Proc. 2016 32nd Picture Coding Symp.*, Dec. 2016, Art. no. 7906379.

[25] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Proc. 8th Int. Conf. Qual. Multimedia Exp.*, 2016.

[26] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF," in *Proc. 2006 IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 1929–1932.

[27] R. Sjoberg *et al.*, "Overview of HEVC high-level syntax and reference picture management," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1858–1870, Dec. 2012.

[28] P. Helle *et al.*, "Block merging for quadtree-based partitioning in HEVC," *IEEE Trans. Circuits Systems Video Technol.*, vol. 22, no. 12, pp. 1720–1731, Dec. 2012.

[29] B. Li, H. Li, L. Li, and J. Zhang, "λ domain rate control algorithm for high efficiency video coding," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3841–3854, Sep. 2014.

[30] L. Li, B. Li, H. Li, and C. W. Chen, "λ domain optimal bit allocation algorithm for high efficiency video coding," *IEEE Trans. Circuits Syst. Video Technol.*, 2017, to be published.

[31] H. Li, B. Li, and J. Xu, "Rate-distortion optimized reference picture management for high efficiency video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1844–1857, Dec. 2012.

[32] Y. Gao, C. Zhu, S. Li, and T. Yang, "Layer-based temporal dependent rate-distortion optimization in random-access hierarchical video coding," in *Proc. 2016 IEEE 18th Int. Workshop Multimedia Signal Process.*, Sep. 2016, pp. 1–6.

[33] K. Andersson *et al.*, "Non-normative HM encoder improvements," Document JCTVC-W0062, San Diego, CA, USA, Feb. 2016.

[34] High Efficiency Video Coding test model, HM-16.7. 2017. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/

[35] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," Document VCEG-M33, Austin, Texas, USA, Apr. 2001.

[36] S. Li, C. Zhu, Y. Gao, Y. Zhou, F. Dufaux, and M.-T. Sun, "Inter-frame dependent rate-distortion optimization using Lagrangian multiplier adaption," in *Proc. 2015 IEEE Int. Conf. Multimedia Expo*, Jun. 2015, pp. 1–6.

[37] S. Li, C. Zhu, Y. Gao, Y. Zhou, F. Dufaux, and M. T. Sun, "Lagrangian multiplier adaptation for rate-distortion optimization with inter-frame dependency," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 117–129, Jan. 2016.



Li Li (M'17) received the B.S. and Ph.D. degrees in electronic engineering from the University of Science and Technology of China, Hefei, China, in 2011 and 2016, respectively.

He is currently a Postdoc Researcher in the University of Missouri—Kansas City, Kansas City, MO, USA. His research interests include image/video coding and processing. He received the Best 10% Paper Award at the 2016 IEEE Visual Communications and Image Processing Conference.



Zhu Li (M'02–SM'07) received the Ph.D. degree in electrical and computer engineering from Northwestern University, Evanston, IL, USA, in 2004.

He is currently an Associate Professor in the Department of Computer Science and Electrical Engineering, University of Missouri—Kansas City, Kansas City, MO, USA, and directs the Multimedia Computing and Communication Lab. He was AFRL Faculty Fellow at the US Air Force Academy in Summer 2016 and 2017; Sr. Staff Researcher/Sr. Manager with Samsung Research America's Multimedia Standards Research Lab, Richardson, TX, 2012–2015; Sr. Staff Researcher/Media Analytics Group Lead with FutureWei (Huawei) Technology's Media Lab, Bridgewater, NJ, 2010–2012; an Assistant Professor in the Department of Computing, The Hong Kong Polytechnic University, from 2008 to 2010; and a Principal Staff Research Engineer with the Multimedia Research Lab, Motorola Labs, from 2000 to 2008. His research interests include image/video processing and compression, machine learning, as well as video adaptation, media network optimization. He holds 25 issued or pending patents, more than 90 publications in book chapters, journals, conference proceedings, and standard contributions in these areas. He became an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA (2015), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEM FOR VIDEO TECHNOLOGY (2016), and the *Journal of Signal Processing Systems* (Springer) in 2015. He is a Steering Committee Member of the IEEE International Conference on Multimedia and Expo (ICME), and an elected member (2014–2017, 2017–2020) of the IEEE Multimedia Signal Processing Tech Committee. Since 2006, he has been the Steering Committee Chair of the IEEE Multimedia Communication Technical Committee. He received the Best Paper Award at the IEEE ICME, Toronto, 2006, and the Best Paper Award at the IEEE International Conference on Image Processing, San Antonio, 2007.



Bin Li (M'14) received the B.S. and Ph.D. degrees in electronic engineering from the University of Science and Technology of China, Hefei, China, in 2008 and 2013, respectively.

He joined Microsoft Research Asia (MSRA), Beijing, China, in 2013 and is currently a Researcher. He has authored or coauthored more than 20 papers. He holds more than ten granted or pending U.S. patents in the area of image and video coding. He has more than 30 technical proposals that have been adopted by Joint Collaborative Team on Video Coding. His current research interests include video coding, processing, and communication.

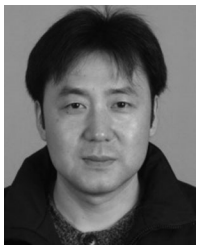
Dr. Li received the Best Paper Award for the International Conference on Mobile and Ubiquitous Multimedia from Association for Computing Machinery in 2011 and the Top 10% Paper Award of 2014 IEEE International Conference on Image Processing. He has been an active contributor to ISO/MPEG and ITU-T video coding standards. He is currently the Co-Chair of the Ad Hoc Group of Screen Content Coding extensions software development.



Dong Liu (M'13) received the B.S. and Ph.D. degrees in electrical engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively.

He joined USTC as an Associate Professor in 2012. He was a Member of Research Staff with the Nokia Research Center, Beijing, China, from 2009 to 2012. He has authored or coauthored more than 30 papers in journals and international conferences. His research interests include image/video coding, multimedia signal processing, and data mining.

Dr. Liu received the 2009 IEEE Transactions on Circuits and Systems for Video Technology Best Paper Award and the Best 10% Paper Award at the 2016 IEEE Visual Communications and Image Processing Conference.



Houqiang Li (M'10–SM'12) received the B.S., M.Eng., and Ph.D. degrees in electronic engineering from the University of Science and Technology of China, Hefei, China, in 1992, 1997, and 2000, respectively.

He is currently a Professor in the Department of Electronic Engineering and Information Science, University of Science and Technology of China. He has authored and coauthored more than 100 papers in journals and conferences. His research interests include video coding and communication, multimedia search, and image/video analysis. He was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2010 to 2013, and has been with the Editorial Board of the *Journal of Multimedia* since 2009. He received the Best Paper Award for Visual Communications and Image Processing in 2012, the Best Paper Award for International Conference on Internet Multimedia Computing and Service in 2012, the Best Paper Award for the International Conference on Mobile and Ubiquitous Multimedia from ACM in 2011, and a senior author of the Best Student Paper of the 5th International Mobile Multimedia Communications Conference (MobiMedia) in 2009.

Light Field Image Coding Using High-Order Intrablock Prediction

Ricardo J. S. Monteiro, *Member, IEEE*, Paulo J. L. Nunes, *Member, IEEE*, Nuno M. M. Rodrigues, *Member, IEEE*, and Sérgio M. M. Faria, *Senior Member, IEEE*

***Abstract*—This paper proposes a two-stage high-order intrablock prediction method for light field image coding. This method exploits the spatial redundancy in lenslet light field images by predicting each image block, through a geometric transformation applied to a region of the causal encoded area. Light field images comprise an array of microimages that are related by complex geometric transformations that cannot be efficiently compensated by state-of-the-art image coding techniques, which are usually based on low-order translational prediction models. The two-stage nature of the proposed method allows us to choose the order of the prediction model most suitable for each block, ranging from pure translations to projective or bilinear transformations, optimized according to an appropriate rate-distortion criterion. The proposed higher order intrablock prediction approach was integrated into a high efficiency video coding (HEVC) codec and evaluated for both unfocused and focused light field camera models, using different resolutions and microlens arrays. Experimental results show consistent bitrate savings, which can go up to 12.62%, when compared to a lower order intrablock prediction solution and 49.82% when compared to HEVC still picture coding.**

***Index Terms*—Geometric transformations, HEVC, high order intra block prediction, light field image coding.**

I. Introduction

LIGHT Field (LF) imaging technology available in lenslet LF cameras allows to jointly capture radiance data and angular information from the light rays hitting the camera's sensor, by means of multiplexing the LF data in a 2D conventional sensor. This is achieved through an array of microlenses, placed between the main lens and the camera sensor. Each microlens creates a micro-image (MI) on the sensor, which is the microlens scene perspective being captured through the main lens. Therefore, a lenslet light field image tends to be like the output of an array of very small cameras.

Manuscript received December 29, 2016; revised April 15, 2017; accepted June 7, 2017. Date of publication June 28, 2017; date of current version October 23, 2017. This work was supported by the Fundação para a Ciência e Tecnologia under Project UID/EEA/50008/2013. The guest editor coordinating the review of this paper and approving it for publication was Prof. Yebin Liu. *(Corresponding author: Ricardo J. S. Monteiro.)*

R. J. S. Monteiro and P. J. L. Nunes are with the Instituto de Telecomunicações, 1049-001 Lisbon, Portugal, and the University Institute of Lisbon, 1649-026 Lisbon, Portugal (e-mail: ricardo.monteiro@lx.it.pt; paulo.nunes@lx.it.pt).

N. M. M. Rodrigues and S. M. M. Faria are with the Instituto de Telecomunicações, 2411-901 Leiria, Portugal, and the Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Leiria, 2411-901 Leiria, Portugal (e-mail: nuno.rodrigues@co.it.pt; sergio.faria@co.it.pt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2721358

The additional knowledge of the scene angular information allows to perform various a posteriori image processing tasks, not straightforwardly possible with traditional cameras. Refocusing and change of perspective after the picture has been taken are the most common examples [1]. These functionalities, derived from the ability to capture the "whole observable" (LF) scene [1], may be advantageous for several applications, like 3D Television [2], since by rendering several views from different perspectives, 2D, 3D and multiview signals can be created; image recognition and medical imaging [3].

Depending on the position of the camera sensor and the microlens array relatively to the main lens, different samplings of the light field can be performed, which define essentially the lenslet LF camera model [4]. Two main types of lenslet LF camera models exist, the unfocused [5] and the focused model [4]. In the classic unfocused camera model case, the sensor is one focal distance away from the microlens array. Thus, the microlens array is focused at infinity, i.e., the light rays that reach the microlens array are parallel [5]. Consequently, the microlens array is completely defocused from the main lens image plane. Therefore, each microlens only captures angular information, meaning that each pixel, within the MI, corresponds to a different angle, or viewpoint [5]. In the focused lenslet LF camera model, the sensor is away from the microlens array focal distance and the microlens array is focused on the main lens image plane, allowing for each microlens to generate a focused MI. This feature allows a higher spatial resolution for rendering, since more than one pixel can be extracted from each MI in the rendering process [4]. These models have been the base for the deployment of this technology, allowing an increasing number of applications and users.

The growing interest in LF technology led the JPEG Committee to launch a new activity, known as JPEG Pleno, to address coding and representation of content generated by emerging imaging technologies such as LF, point-cloud and holographic technologies [6].

The large amount of data required to adequately represent a LF scene, when compared to the case of typical 2D pictures, calls for efficient techniques for both transmission and storage of this type of content. In this context, several authors proposed specific LF coding techniques, which can be applied directly to the lenslet LF images, in order to exploit the MIs redundancy. Alternatively, other techniques are applied to a different representation of the same LF, which comprises the view point

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

images, also known as sub-aperture images (SAIs). The SAIs are generated by extracting at least one pixel, in a fixed position, from each MI and organizing them into a matrix. Each SAI represents a rendered image, from a different perspective, extracted from the LF image.

State-of-the-art LF coding schemes rely on block matching techniques to exploit the inherent spatial redundancy in lenslet LF images. However, these low order prediction (LOP) models use only two degrees of freedom (DoF), as only translations are used to describe the inherent LF image spatial redundancy.

Due to the small baseline between MIs in lenslet LF images, the different MIs can be approximately related by changes in perspective, which require eight DoF to be described. To the best of authors' knowledge, this characteristic has not been exploited by previous LF coding approaches described in the literature. This additional matching accuracy is important to develop a coding method able to cope with important features of the LF content, such as:

1) *The LF camera model,* i.e.*,* both focused and unfocused models should be handled;
2) *The type of microlens array structure,* e.g., rectangular or hexagonal microlens layouts, creating rectangular, hexagonal or circular MIs;
3) *The MI size*, i.e., a parameter that depends on the camera, and has a strong influence on the number of possible rendered points of view and their spatial resolution.

High order prediction (HOP) models, e.g., using geometric transformations with more DoF, have been studied during the last two decades in traditional 2D and 3D image coding scenarios. Several geometric models, like translation, rotation, scale, shear and perspective changes have been used to improve the coding efficiency, by exploiting spatial [7], temporal [8]–[13] and inter-view [14]–[17] redundancy. In most proposals, these models have been applied image-wise (instead of block-wise), due to two main reasons: (i) high computational complexity in block-wise model parameter estimation, and (ii) significant additional bit rate required for parameter transmission. Despite these drawbacks, this paper demonstrates that block-wise HOP models can increase block matching accuracy and, thus, coding efficiency for lenslet LF images.

The method proposed in this paper for encoding lenslet LF images relies on a two-stage block-wise HOP model, where each image block is intra predicted from a reference in the causal area of the image, i.e., containing pixels that were already encoded. Since this approach is applied block-wise, it is possible to optimize the HOP model (number of DoF) for each block to be encoded. Taking advantage of the extra DoF available in HOP models, it is possible to outperform state-of-the-art coding techniques based on LOP models.

The remainder of this paper is organized as follows: Section II presents a review of several relevant state-of-the-art solutions, regarding LF image coding; Section III describes the geometric transformations used in the proposed prediction method; Section IV presents the proposed HOP model; Section V presents the test conditions and experimental results; and, finally, Section VI concludes the paper.

II. Related Work on Light Field Image Coding

Several schemes to encode lenslet LF images are described in the literature, aiming to exploit the intra-LF image redundancy. These schemes rely on different LF image representations and coding techniques, which may be categorized according to the fundamental adopted approach as: transform-based coding, pseudo-video sequence coding, disparity-based coding and non-local spatial prediction coding.

A. Transform-Based Coding

Some LF coding schemes rely, essentially, on the use of a transform, mainly the discrete cosine transform (DCT) [18], [19] or the discrete wavelet transform (DWT) [20]. In [18], a 3D-DCT is applied to a stack of MIs, to exploit the existing spatial redundancy within a MI, as well as the redundancy between adjacent MIs. In [20], a LF image is decomposed into SAIs, and a 3D-DWT is applied to a stack of these SAIs. The lower frequency bands are transformed using a two-dimensional discrete wavelet transform (2D-DWT), while the remaining higher frequency coefficients are simply quantized and arithmetic encoded. These coding schemes are reportedly more efficient than JPEG, but not as efficient as high efficiency video coding (HEVC) still picture coding.

B. Pseudo-Video Sequence Coding

This type of LF coding schemes represent the LF image as a set of MIs or SAIs, and re-organize them into a low resolution pseudo-video sequence (PVS), which is then compressed using a standard video encoder. Various scanning strategies to order the PVS are considered to better exploit the redundancy between MIs or SAIs. Dai *et al*. [21] propose to scan the SAIs using either a raster or a spiral scan and then encode the generated video sequence with H.264/AVC. Vieira *et al*. [22] used similar scanning strategy combined with several prediction structures supported by HEVC. In both cases it is possible to conclude that the spiral scan is more efficient than the raster scan. More recently, in the ICME light field compression challenge [23], Liu *et al*. [24] used a PVS scheme to organize the SAIs into layers, depending on the proximity to the central view, starting with the central SAI and moving on to the outer views. The more distant the SAI is from the center, the higher the value of the used quantization parameter (QP) should be. This scheme was implemented using both HEVC test model (HM) and JEM [25] software. Because the rate allocation is not uniform along the LF image, this method is prone to reconstruct views with different objective qualities.

C. Disparity-Based Coding

In this type of LF coding schemes the LF image is considered as a set of views captured by different cameras (either in the form of MIs or SAIs), which may be encoded exploiting inter-view disparity. In [26], the authors propose a coding method that uses some SAIs to calculate a set of disparity maps prior to coding, which are then used to predict the remaining SAIs. The authors concluded that this approach is suitable to encode

synthetic images, where disparity compensation alone can be enough to predict a SAI. A compression scheme that incorporates disparity compensation into 4D wavelet coding using disparity compensated lifting is proposed in [27]. The disparity information derived from an approximated model of the scene is applied to modify the update and prediction filters of the lifting procedure. In [28], the authors propose a scalable (two-layer) LF coding approach for the focused LF camera model, using a LF representation that consists of a sparse set of MIs and associated disparity maps. Based on the sparse set of MIs and the associated disparity maps (first layer), a reference prediction LF image is obtained through a reconstruction method that relies on disparity-based interpolation and inpainting. This reconstructed LF image is then used to encode the original LF image (second layer), by encoding the prediction residue. This approach was later extended [29] with a third layer of scalability and the use of lossy encoded disparity maps, in contrast with the lossless transmission of the disparity maps, used in the first approach. Both versions of the work are able to outperform HEVC still picture coding.

D. Non-Local Spatial Prediction Coding

Several methods to exploit the non-local spatial redundancy were proposed as additional coding tools for existing video coding standards, like HEVC. In [30], a self-similarity compensated prediction is proposed to take advantage of the flexible partition patterns used by this video codec. In [31] this method was extended with a bi-directional mode to increase its coding efficiency. Additionally, in [32], an alternative non-local spatial prediction method has been investigated, relying on a prediction mode based on locally linear embedding integrated in HEVC. Differently from the other schemes, that exploit non-local spatial redundancy, this method distributes the computational complexity between the encoder and the decoder, i.e., the locally linear embedding procedure must be replicated in the decoder. In [33] the authors developed a multi-hypothesis coding method specifically for focused LF image and video. This method uses up to two hypotheses for prediction in both spatial and time domains, which outperforms single-hypothesis based prediction. For the unfocused camera model, the authors concluded that the rate-distortion efficiency is still much higher, compared to JPEG or HEVC, however the gains relatively to HEVC are smaller in this case when compared to the focused model [34].

The main advantage of this category is that, in most approaches, the lenslet LF images are encoded without the need of any pre-processing steps or any prior knowledge about the capturing device, e.g., the LF camera model, the microlens array structure and the MI size.

III. Geometric Transformations for High Order Prediction

In most state-of-the-art encoders, prediction between blocks of pixels is performed using very simple transformations, like translations. However, a lenslet LF image is comprised of MIs that are related by more complex transformations, resulting from the fact that each MI represents the scene being captured from slightly different perspectives. In such cases, it is advantageous to use geometric transformations that better exploit the features of the LF image and its MIs.

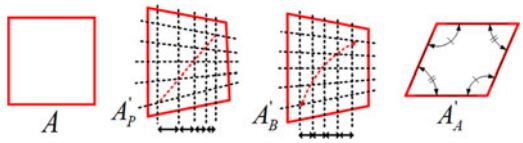


Fig. 1. Examples of possible GTs applied to block A: Projective (A'_P), bilinear (A'_B) and affine (A'_A).

A geometric transformation (GT) is able to map perspective changes from one view (generically associated to a quadrilateral) into another view, requiring up to eight DoF.

Considering two different blocks, A and A', each one with its own coordinate system, (u, v) and (x, y), respectively, it is possible to define a generic relationship:

$$(x, y) = (X(u, v), Y(u, v)), \tag{1}$$

where X and Y are mapping functions for each coordinate. These functions create a point to point correspondence between images. Depending on the number of DoF used by the mapping functions in (1), different number of independent point to point correspondences are possible. To describe these mapping functions, some GTs may be used, namely, Projective, Bilinear or a simpler Affine GT, as illustrated in Fig. 1.

A. Projective Geometric Transformation

In order to simplify the mathematics used in this kind of GT, homogeneous coordinates are commonly used [35]. Thus, the Projective GT can be defined by a 3×3 matrix $\boldsymbol{H}$ verifying (2):

$$[x, y, 1] = [uh, vh, h]\,\boldsymbol{H}. \tag{2}$$

The Projective matrix $\boldsymbol{H}$ can be decomposed into three different submatrices, $\boldsymbol{L_p}$, $\boldsymbol{T_p}$ and $\boldsymbol{P_p}$:

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{L_p} & \boldsymbol{P_p} \\ \boldsymbol{T_p} & 1 \end{bmatrix}$$
$$\boldsymbol{L_p} = \begin{bmatrix} l_{00} & l_{01} \\ l_{10} & l_{11} \end{bmatrix},\ \boldsymbol{T_p} = \begin{bmatrix} t_x & t_y \end{bmatrix},\ \boldsymbol{P_p^T} = \begin{bmatrix} p_x & tp_y \end{bmatrix} \tag{3}$$

Each submatrix is responsible for a different elementary type of GT: $\boldsymbol{T_p}$ is responsible for the description of translations, $\boldsymbol{L_p}$ is able to define linear transformations such as rotation, scaling, and shearing, and $\boldsymbol{P_p}$ describes perspective transformations.

To fully exploit the capabilities of the projective matrix $\boldsymbol{H}$, a four-point correspondence is necessary between blocks A and A'. In this case, the full transformation matrix corresponds to the following system of equations:

$$\begin{cases} x = X(u, v) = \dfrac{l_{00}u + l_{10}v + t_x}{p_x u + p_y v + 1} \\ y = Y(u, v) = \dfrac{l_{01}u + l_{11}v + t_y}{p_x u + p_y v + 1} \end{cases} \tag{4}$$

The system of equations (4) defines the necessary calculations for mapping the coordinates of every pixel of block *A* into the transformed block A'.

The number of available DoF is directly related with the number of known points of correspondence which exist between both images. For less than four points of correspondence, simpler transformations can be represented by the perspective model. For example, if one point is known, the only component that can be possibly described is a translation, i.e., $\boldsymbol{T_p} = [t_x, t_y]$, $\boldsymbol{P_p} = [0,0]^T$ and $\boldsymbol{L_p} = \boldsymbol{I}$. This case is defined by (5):

$$[x, y, 1] = [uh, vh, h] \begin{bmatrix} \boldsymbol{I} & 0 \\ \boldsymbol{T_p} & 1 \end{bmatrix}, \tag{5}$$

which can be translated into the system of equations (6):

$$\begin{cases} x = X(u) = u + t_x \\ y = Y(v) = v + t_y \end{cases}. \tag{6}$$

B. Bilinear Geometric Transformation

The Bilinear GT is an alternative to the Projective GT, defined by a 4×2 matrix $\boldsymbol{B}$ verifying (7):

$$[x, y] = [uv, u, v, 1]\, \boldsymbol{B}, \tag{7}$$

where:

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{P_b} \\ \boldsymbol{L_b} \\ \boldsymbol{T_b} \end{bmatrix} = \begin{bmatrix} p_x & p_y \\ l_{00} & l_{01} \\ l_{10} & l_{11} \\ t_x & t_y \end{bmatrix}, \tag{8}$$

The Bilinear GT matrix $\boldsymbol{B}$ can represent similar GTs as the Projective GT, with the same number of DoF. However, it performs a non-planar transformation, which makes it more flexible. Thus, only horizontal and vertical lines, as well as equispaced points along these directions, are preserved [36]. Diagonal lines, on the other hand, are not mapped as lines but as quadratic curves. This feature is illustrated in Fig. 1, where, in the case of the Bilinear GT, points along vertical parallel lines are kept equispaced, while the points along diagonal lines are mapped onto a quadratic curve (block A_B). When the Projective GT (block A'_P) is used, points along the parallel vertical lines do not stay equispaced but points along diagonal lines are also mapped along a line. Another property of this GT, when compared to the Projective GT, is the need for simpler calculations per pixel, given by (9):

$$\begin{cases} x = X(u, v) = uvp_x + ul_{00} + vl_{10} + t_x \\ y = Y(u, v) = uvp_y + ul_{01} + vl_{11} + t_y \end{cases}. \tag{9}$$

C. Affine Geometric Transformation

When using either the Projective or the Bilinear GT, eight DoF are available. However, a simpler case exists, which is known as the Affine GT, that is able to describe GTs up to six DoF. The Affine GT can be described as a particular case of Projective or Bilinear GTs, by using matrices $\boldsymbol{H}$ and $\boldsymbol{B}$ with $\boldsymbol{P_p^T} = [0 \;\; 0]$ and $\boldsymbol{P_b} = [0 \;\; 0]$, respectively. This GT only requires three points of correspondence between images, defined by (10):

$$\begin{cases} x = X(u, v) = l_{00}u + l_{10}v + t_x \\ y = Y(u, v) = l_{01}u + l_{11}v + t_y \end{cases}. \tag{10}$$

IV. Proposed High Order Prediction Mode

This section proposes a LF image coding method, based on a high order prediction model, which is implemented as a block-wise prediction mode in HEVC. This HOP mode is added to the set of HEVC Intra prediction modes, i.e., Planar mode, DC mode and the 33 intra Directional modes.

The proposed HOP mode predicts each block by applying a GT between two quadrilaterals, the current block and a block in the reference region, the causal area of pixels already encoded. The algorithm for the proposed prediction mode can be described through the following steps:

1) *Selection of the next set of correspondence points to be evaluated*: Selection of a quadrilateral in the causal area of pixels (from a set of pre-defined cases), with corners $\{P'_n\}$, that is mapped into the block which is being predicted, with corners $\{P_n\}$ (see left side of Fig. 2);
2) *Calculation of the GT parameters*: Calculation of the transformation parameters that map the quadrilateral defined by $\{P'_n\}$ into the one defined by $\{P_n\}$;
3) *Inverse GT mapping*: Mapping of the causal quadrilateral defined by $\{P'_n\}$ to the one defined by $\{P_n\}$, using an inverse mapping procedure with the parameters calculated in the previous step, in order to compute the block prediction error; error and the estimated number of bits to transmit the GT parameters;
4) *Estimation of the GT RD cost*: Estimation of the rate-distortion (RD) cost, *J*, associated to the GT that is being evaluated, considering the computed block prediction;
5) *Repeat the above steps to find the GT with minimum RD cost:* Evaluate iteratively all the pre-defined combinations of correspondence points and choose the one that has the minimum RD cost J.
6) *Encode the HOP mode information:* The corner displacement between the quadrilateral in the causal area $\{P'_n\}$ and the block which is being predicted $\{P_n\}$ is signaled to the decoder.

The following sub-sections explain each step of the proposed HOP mode in more detail.

A. Selection of the Correspondence Points

The major challenges faced by the proposed algorithm are the computational complexity required to estimate the optimal set of GT parameters and the necessary bit-rate for transmitting this data. To tackle both problems, a rate-distortion-complexity tradeoff is defined. From Fig. 2 (left side) it can be inferred that, if all possible four-point correspondences between the prediction block and the current block to be encoded were evaluated, the number of tested transformations per block would be larger than $(2\,W^2)^4$, i.e., for a search window ($W = 128$) more than 1.15×10^{18} correspondence possibilities per block exist. To

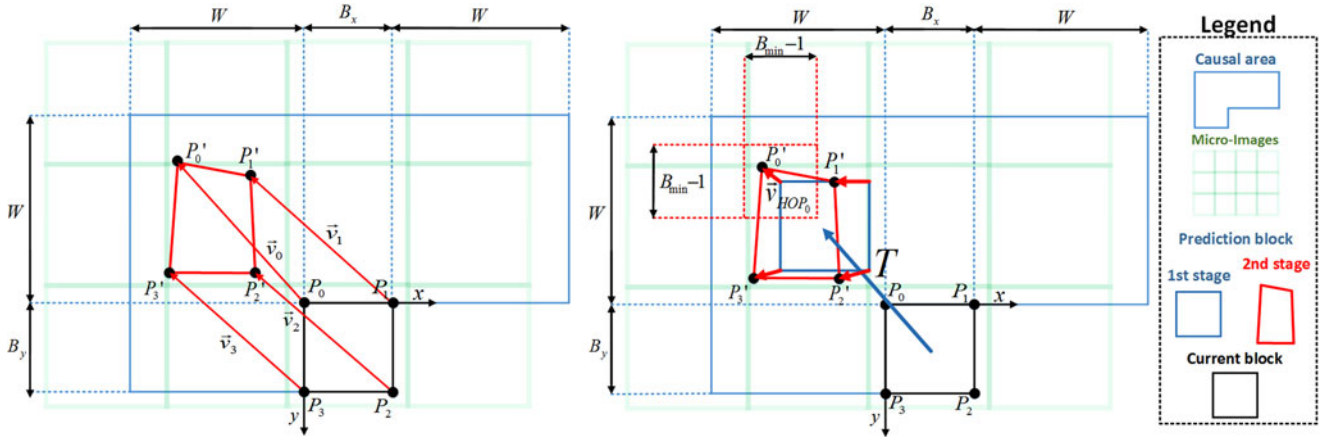


Fig. 2. Block prediction using a HOP model: generic single-stage HOP model mapping (left side), and proposed two-stage HOP model mapping (right side).

reduce the number of tests to a practicable number, a two-stage minimization problem is proposed, aiming to determine a good approximation to the optimal HOP model, as illustrated in Fig. 2 (right side):

1) LOP Model Estimation: In the first stage, a pure translational LOP model (two DoF) is used. The result of this stage is, the bidimensional vector, $\boldsymbol{T}$, with the lowest RD cost, pointing into the search window of the causal area (see the blue vector on the right side of Fig. 2). The search to determine $\boldsymbol{T}$ is performed using a full search algorithm, as described in [30]. The LOP estimation stage of the proposed HOP mode is based on the Self Similarity (SS) prediction method. The prediction cost is minimized by testing all the possible positions inside the search window for a single vector that relates the current block to the prediction block. The $\boldsymbol{T}$ vectors, generated by the first stage, can be either encoded explicitly, similarly to motion vectors in HEVC or using the SS-Skip mode, which creates a list of candidates that includes the $\boldsymbol{T}$ vectors used to encode neighboring blocks. If one candidate from this list is selected to encode the current block, it is only necessary to encode its index, as in the HEVC merge mode. Additionally, some predetermined vectors are added to the candidate list, referred to as MI-based candidates [30]. These candidates correspond to vectors that are very likely to be selected by the SS prediction mode, such as, vectors pointing to the same spatial position of the current MI within the left, above and above-left MIs.

2) HOP Model Estimation: In the second stage, a HOP model (up to eight DoF) is used, employing as a starting point the result of the first stage (see, respectively, the red and blue quadrilaterals on the right side of Fig. 2). For this, a set of four vectors, $\{v_{HOP_n}\}$, is computed, each of them defining the position of one corner of the reference quadrilateral, thus defining the 2D GT.

To further reduce the computational complexity of the second stage of this minimization problem, a 2D logarithmic fast search method has been adopted, which is applied to each corner of the prediction block (blue rectangle). In this case, the maximum number of search steps has been set to $\log_2(\min(B_x,\ B_y)) - 1$, depending on the size of the prediction block, i.e., B_x (width) and B_y (height). In each step, the searching points are defined

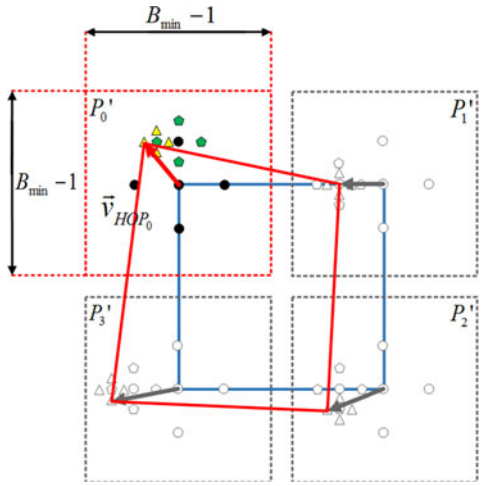


Fig. 3. Fast search method adopted for each corner of the prediction block (blue rectangle) used to estimate the HOP model (red quadrilateral).

according to a five-point small diamond-shaped basis pattern with an initial search step size equal to $\min(B_x,\ B_y)/4$ [37], [38]. This 2D logarithmic fast search method using the five-point small diamond-shaped basis pattern is graphically represented in Fig. 3 across three search steps, represented, respectively, by black circles, green pentagons and yellow triangles. After each search step, the point that minimizes the RD cost function is set as the center of the next step and the search step size is halved until a unitary step value is reached. In the example of Fig. 3, in the first corner (P_0), the five points associated with the first step, represented by the black circles, are tested. The point that minimizes the RD cost function for the first search step is the black circle on the top. For the second and third search steps, the points on left, respectively, green pentagon and yellow triangle, are the points that yield the lowest RD cost. The final point is selected to define the red arrow that describes the corner displacement of the first corner of the block.

Considering that the search procedure must be applied to all the corners of the prediction block over several search steps, there are two ways of implementing this second stage search: by jointly optimizing each step of the search procedure

for the four corners or by independently optimizing each step of the search procedure. By considering five points for each of the S search steps of the 2D logarithm search, the required number of search points for each option is given by $(5 \times S)^n$ or $5^n \times S$, where n is the number of corners. In order to reduce the computational complexity, the second option was used, where each step is optimized individually.

The stop condition for this search method is met when the corner step size reaches the unit. Therefore, the example shown in Fig. 3 represents the unitary steps as the yellow triangles. Since the underlying codec uses variable block sizes, S will depend on the block size. The search window for each corner is limited to $\min(B_x, B_y) - 1$, as illustrated in Fig. 3 (see the dashed red block).

The quadrilateral used by the HOP model estimation may be scaled to increase pixel precision. Figs. 2 and 3 illustrate the second stage applied to a blue rectangle with the same size of the block being predicted (in black) to not overload the figures. However, in our implementation, a rectangle, twice the size of the original block, is used to determine the HOP model. This modification means that an integer pixel displacement in one of the corners of the large quadrilateral corresponds to a sub-pixel displacement in the area of the original rectangle. For a block twice the size of the original block, one extra search step is performed by the 2D logarithmic search algorithm that is used at the HOP stage. This occurs because the stopping condition for the search algorithm is the unitary step size. The pixel precision can be further extended by using a rectangle with sides four or eight times the size of the original blue rectangle, which increase, the number of search steps by one or two, respectively. After extensive testing, the best solution in a RD sense was adopted, that is increasing the blue rectangle to twice the original size, despite requiring one extra step.

As the second stage of the HOP search can be biased by the first stage result, the global result of the search method also tests the $\boldsymbol{T}$ vectors used in the previously encoded neighboring blocks (vector predictors), instead of considering only the best $\boldsymbol{T}$ vector from the first stage. Additionally, other $\boldsymbol{T}$ vectors can be tested in conjunction with the HOP model estimation, e.g. top ten candidates from the first stage. However, it was experimentally verified by the authors that the vectors that are more RD cost efficient are the $\boldsymbol{T}$ vector predictors.

The proposed approach can be implemented using either the Projective GT defined in (3):

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{T} & 0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{L'_p} & \boldsymbol{P'_p} \\ \boldsymbol{T'_p} & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{L'_p} & \boldsymbol{P'_p} \\ \boldsymbol{T} + \boldsymbol{T'_p} & 1 \end{bmatrix} \quad (11)$$

or the Bilinear GT defined in (8):

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{T} \end{bmatrix} + \begin{bmatrix} \boldsymbol{P'_b} \\ \boldsymbol{L'_b} \\ \boldsymbol{T'_b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{P'_b} \\ \boldsymbol{L'_b} \\ \boldsymbol{T} + \boldsymbol{T'_b} \end{bmatrix}. \quad (12)$$

Where $\boldsymbol{T}$ is the vector estimated during the LOP stage and $\boldsymbol{T'}$, $\boldsymbol{L'}$ and $\boldsymbol{P'}$ are the GT parameters that describe the HOP stage.

B. Calculation of the GT Parameters

After obtaining vector $\boldsymbol{T}$ (see the right side of Fig. 2), it is possible to determine submatrices $\boldsymbol{T'}$, $\boldsymbol{P'}$ and $\boldsymbol{L'}$ in (11) and (12), by using their width and height, B_x and B_y, respectively, and the small vectors associated with the corner position change of the blue rectangle:

$$\begin{cases} \vec{v}_{HOP_0} = (u_0, v_0) \\ \vec{v}_{HOP_1} = (u_1 - (B_x - 1), v_1) \\ \vec{v}_{HOP_2} = (u_2 - (B_x - 1), v_2 - (B_y - 1)) \\ \vec{v}_{HOP_3} = (u_3, v_3 - (B_y - 1)) \end{cases} \quad (13)$$

Note that in the proposed two-stage approach, vectors $\vec{v}_n$, represented in the left of Fig. 2, correspond to the sum of vector $\boldsymbol{T}$ from the first stage, with the four smaller vectors from the second stage, i.e., $\vec{v}_n = T + \vec{v}_{HOP_n}$, represented in the right of Fig. 2. If the Projective GT is used some auxiliary variables are defined:

$$\begin{cases} \Delta u_1 = u_1 - u_2 \\ \Delta u_2 = u_3 - u_2 \\ \Delta u_3 = u_0 - u_1 + u_2 - u_3 \end{cases}$$

$$\begin{cases} \Delta v_1 = v_1 - v_2 \\ \Delta v_2 = v_3 - v_2 \\ \Delta v_3 = v_0 - v_1 + v_2 - v_3 \end{cases}. \quad (14)$$

The Affine GT can be defined by any three of the four vectors ($\vec{v}_{HOP}$). In this paper, the first three vectors,$\vec{v}_{HOP_0}$,$\vec{v}_{HOP_1}$ and $\vec{v}_{HOP_2}$ are generated using the second stage of the proposed approach, where the remaining vector, $\vec{v}_{HOP_3}$, is calculated assuming $\Delta u_3 = \Delta v_3 = 0$, thus resulting in $\vec{v}_{HOP_3} = (u_0 - u_1 + u_2, v_0 - v_1 + v_2 - (B_y - 1))$.Using (14), the individual parameters in the submatrices can then be calculated by (15) for the Projective GT:

$$\boldsymbol{P'_p} = \begin{bmatrix} & \begin{vmatrix} \Delta u_3 & \Delta u_2 \\ \Delta v_3 & \Delta v_2 \end{vmatrix} \\ \frac{1}{B_x - 1} & \begin{vmatrix} \Delta u_1 & \Delta u_2 \\ \Delta v_1 & \Delta v_2 \end{vmatrix} \\ & \begin{vmatrix} \Delta u_1 & \Delta u_3 \\ \Delta v_1 & \Delta v_3 \end{vmatrix} \\ \frac{1}{B_y - 1} & \begin{vmatrix} \Delta u_1 & \Delta u_2 \\ \Delta v_1 & \Delta v_2 \end{vmatrix} \end{bmatrix} \text{ and}$$

$$\boldsymbol{L'_p} = \begin{bmatrix} \frac{u_1 - u_0}{B_x - 1} + p_x u_1 & \frac{u_3 - u_0}{B_y - 1} + p_y u_3 \\ \frac{v_1 - v_0}{B_x - 1} + p_x v_1 & \frac{v_3 - v_0}{B_y - 1} + p_y v_3 \end{bmatrix}. \quad (15)$$

Similarly, for the Bilinear GT, the corresponding submatrices are calculated by (16):

$$\boldsymbol{P'^T_b} = \begin{bmatrix} \frac{u_0 - u_1 + u_2 - u_3}{(B_x - 1)(B_y - 1)} \\ \frac{v_0 - v_1 + v_2 - v_3}{(B_x - 1)(B_y - 1)} \end{bmatrix} \text{ and}$$

$$\boldsymbol{L'_b} = \begin{bmatrix} \frac{u_3 - u_0}{B_y - 1} & \frac{v_3 - v_0}{B_y - 1} \\ \frac{u_1 - u_0}{B_x - 1} & \frac{v_1 - v_0}{B_x - 1} \end{bmatrix}. \quad (16)$$

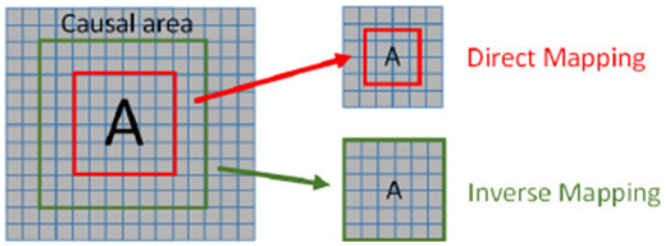


Fig. 4. Example of direct mapping and inverse mapping when a scale GT is applied.

For both cases we have:

$$\boldsymbol{T}'_{\boldsymbol{p}} = \boldsymbol{T}'_{\boldsymbol{b}} = \begin{bmatrix} u_0 & v_0 \end{bmatrix}. \tag{17}$$

C. Inverse GT Mapping

As previously mentioned, a GT between two blocks corresponds to a mapping of every pixel within one block into the other block, e.g., the mapping functions (4) and (9) correspond to the Projective and Bilinear GT, respectively. When the mapping is performed from the rectangular block to be encoded to an arbitrary reference quadrilateral it is called a direct mapping, otherwise it is called an inverse mapping. An example of how both mapping procedures for a simple scaling GT can be found in Fig. 4. As can be observed in Fig. 4, when direct mapping is used the final quadrilateral shape (red block) does not match the desired reference block pixel grid, requiring to perform pixel interpolation prior to calculate the distortion between the transformed block and the reference block. For the sake of simplicity, an inverse mapping has been adopted, as it generates a rectangular prediction block with the same dimensions of the block to be encoded.

Thus, regardless of the size of the quadrilateral used for estimation, (4) and (9) take as input the coordinates of the block to be encoded, i.e., $u \in [0, B_x - 1]$ and $v \in [0, B_y - 1]$, and generate as output the coordinates (x, y) in the causal area, where the reference pixel value is going to be extracted from. Since x and y are typically fractional values, a bilinear interpolation filter is used to compute the actual pixel value.

D. Estimation of the GT RD Cost

The optimal HOP model for each block is determined through RD optimization, minimizing the associated Lagrangian cost, $J = D + \lambda R$, over the entire set of pre-defined GT. D refers to the distortion between the prediction block and the current block, R is the estimated number of bits used to encode the block using the GT under evaluation, and λ is the Lagrange multiplier, computed as in HM version 15.0 for Intra-coded frames. The parameter λ is the same for all prediction modes, including the intra modes, so no biases in terms of prediction mode selection are introduced. In this paper, D, is computed as the sum of absolute differences (SAD) in the pixel domain, in the first stage, and SAD in the Hadamard domain, in the second stage, as suggested in [39].

By using a two-stage method it is possible to evaluate if it is more advantageous to use LOP or HOP for each block, by comparing the associated costs, given by:

$$J_{LOP} = D_{LOP} + \lambda R_{LOP} \text{ and}$$
$$J_{HOP} = D_{HOP} + \lambda R_{HOP}, \tag{18}$$

where R_{LOP} and R_{HOP} are the estimated number of bits for the corresponding coding mode.

The usage of LOP or HOP is conveyed to the decoder through a binary flag, F_{HOP}. When LOP is considered more efficient in a RD sense, $F_{HOP} = 0$, and only $\boldsymbol{T}$ is transmitted in the bitstream. On the contrary, if HOP is used, all the elements that describe $\boldsymbol{T}$ are transmitted, followed by $F_{HOP} = 1$ and the four additional $\vec{v}_{HOP_n}$ vectors.

The number of bits required to signal the HOP mode, R_{HOP}, is the sum of R_{LOP} and the estimated bits for encoding the four vectors, $\vec{v}_{HOP_n}$, that define the used HOP model. The rate of these small amplitude vectors is estimated using the same procedure as vector $\boldsymbol{T}$.

E. Encode the HOP Mode Information

After finding the optimal HOP model, the cost of the HOP mode, J_{HOP}, is compared against the cost of the other intra prediction modes, i.e., DC, Planar and the 33 Directional modes, and the mode with the lowest RD cost is encoded. For this, the context adaptive binary arithmetic coding (CABAC) entropy coding method used by HEVC is used to encode the HOP mode information. The CABAC entropy coder is based on three steps: (i) binarization of syntax elements, (ii) context modeling, and (iii) binary arithmetic coding. In this implementation, these three steps have been maintained using, however, new contexts. Vectors $\boldsymbol{T}$ and $\vec{v}_{HOP_n}$, and lag, F_{HOP}, are transmitted to the decoder using the HM approach for motion vectors and merge flags [40].

To encode $\boldsymbol{T}$, the same syntax elements of HEVC for motion data are used, i.e., motion vector differences, MVP index, reference picture list (RPL) and RPL index.

The way the HOP model information is conveyed to the decoder can highly influence the coding efficiency. One possible approach is to send the GT parameters, i.e., in the $\boldsymbol{H}$ or $\boldsymbol{B}$ matrix, which need to be represented with high precision. Alternatively, as proposed in this paper, the encoder just sends the four vectors, $\vec{v}_{HOP_n}$, which can be represented with just a few bits. The major advantage of encoding the GT parameters matrix is that they do not need to be recalculated at the decoder side through (13)–(17). However, they need to be encoded with a very high precision because these values are not very robust to quantization [11]. Consequently, encoding the vectors, $\vec{\boldsymbol{v}}_{\boldsymbol{HOP_n}}$, leads to higher compression efficiency.

V. Experimental Results

In this section the performance of the proposed lenslet LF coding solution, incorporating the HOP mode, is evaluated in comparison with state-of-the-art coding solutions based on LOP approaches. First, this section describes the test conditions, including the used lenslet LF test images, the benchmark solutions and the relevant test parameters. Afterwards, experimental results comparing the RD performance of different types



Fig. 5. LF test images part of the experimental test setup. First row (from left to right): Plane and Toy (frame 0 and 150), Demichelis Spark (frame 0), Demichelis Cut (frame 0), Laura and Seagull. Second and third rows: sub-set of the LF EPFL dataset.

of prediction models are presented and discussed. These results are complemented with some statistical information about prediction mode usage and an evaluation of the quality of the rendered views, as proposed in [23], using the coded LF images.

A. Test Conditions

In order to evaluate the RD performance of the proposed LF coding solution, two types of LF images were selected for the experimental test setup. The first type of images were acquired using LF cameras with a focused (FOC) optical setup [41], [42]. The second type of images were acquired using a Lytro Illum camera that is commercially available and uses an unfocused (UNF) optical setup. This second set of images constitutes the dataset used for the 2016 ICME Grand Challenge on LF image compression extracted from the EPFL dataset [23]. The central rendered views of all the test images are shown in Fig. 5, where the first row corresponds to the first type of images and the second and third rows correspond to the second one. This selection includes LF images with different resolutions, MI resolutions and types of microlens arrays, with different MI shape. *Plane and Toy* images have a resolution of 1920×1088 (MIs 28×28); *Demichelis* images have a resolution of 2880×1620 (MIs 38×38); *Laura* and *Seagull* have a resolution of 7240×5432 (MIs 75×75); EPFL images have a resolution of 7728×5368 pixels (MIs 15×15).

The proposed HOP mode was implemented into the HEVC test model version 15.0 (HM 15.0) as an additional intra prediction mode. This LF codec, corresponding to the proposed solution, will be referred to as HEVC-HOP, where HEVC using only the standard Intra modes is simply referred to as HEVC. Additionally, the work in [30] is used as benchmark for RD performance and it is referred to as HEVC-SS.

The common HM test conditions were adopted, using QP values of 22, 27, 32 and 37. The causal window size W is 128 for both HEVC-HOP and HEVC-SS, for every encoded image. The number of available SS or $\boldsymbol{T}$ vector predictors, used for coding, is 2. These $\boldsymbol{T}$ vector predictors are used as additional vector $\boldsymbol{T}$ candidates for the LOP model estimation stage. As mentioned in the previous section, these alternative vectors are tested in order to have a more unbiased result when estimating the HOP model. The number of candidates available for SS-Skip is 5 in both HEVC-SS and HEVC-HOP.

B. Experimental Results

All the LF images in Fig. 5 are encoded and decoded using the HEVC, HEVC-HOP and HEVC-SS codecs, and the RD performance is evaluated using a Bjøntegaard Delta Metric. Additionally, several variants of HEVC-HOP are tested. These variants of HEVC-HOP are referred to as HEVC-HOP-A, HEVC-HOP-P and HEVC-HOP-B, respectively for Affine (six DoF), Projective (eight DoF) and Bilinear (eight DoF) GTs. Table I shows the RD performance comparison between HEVC and HEVC-SS, and between HEVC-SS and the various HEVC-HOP variants.

1) Comparison Between LOP and HOP: Table I shows that HEVC-SS can outperform HEVC, for all tests, with bitrate savings up to 45.35%. Nevertheless, all versions of the proposed HEVC-HOP method are even more efficient than HEVC-SS to encode LF images. This increased performance, with bitrate savings up to 12.62% for certain images relatively to HEVC-SS (49.82% relatively to HEVC), comes from the use of a higher order prediction model. Since HEVC-SS is limited to two DoF, it is not able to accurately describe block transformations more complex than a simple translation. When comparing the results by means of comparing the effectiveness of adding prediction tools with more than two DoF, it is possible to notice that for the encoded LF images, the best case is when eight DoF are used. If eight DoF are available, i.e., when HEVC-HOP-P is being used, four points of correspondence are transmitted, which allows the description of not only translations, but also rotations, scaling, shearing and perspective changes. In this case, although extra information needs to be encoded, relative to the HEVC-SS case, the bitrate savings increases to 5.86% (28.81% relative to HEVC), in average, for all tested LF images.

2) Comparison Between the Proposed GTs: The proposed prediction mode HEVC-HOP-B, using a Bilinear GT, can achieve similar results to HEVC-HOP-P for most images, both

TABLE I
BD-PSNR-Y AND BD-RATE RESULTS COMPARING HEVC, HEVC-SS (TWO DOF) AND HEVC-HOP, USING SIX DOF AND EIGHT DOF AND TWO DIFFERENT KINDS OF GEOMETRIC TRANSFORMATIONS

Image	HEVC-SS (2 DoF) vs HEVC		HEVC-HOP-A (6 DoF) vs HEVC-SS (2 DoF)		HEVC-HOP-P (8 DoF) vs HEVC-SS (2 DoF)		HEVC-HOP-B (8 DoF) vs HEVC-SS (2 DoF)	
	BD-PSNR-Y	BD-RATE	BD-PSNR-Y	BD-RATE	BD-PSNR-Y	BD-RATE	BD-PSNR-Y	BD-RATE
Plane and Toy, frame 0 (PT0)	0.90 dB	−14.64%	0.23 dB	−3.93%	**0.27 dB**	**−4.66%**	0.23 dB	−4.02%
Plane and Toy, frame 150 (PT150)	1.44 dB	−19.02%	0.64 dB	−9.44%	**0.75 dB**	**−11.05%**	0.71 dB	−10.50%
Demichelis Spark, frame 0 (DS)	1.09 dB	−31.43%	0.23 dB	−7.41%	**0.26 dB**	**−8.39%**	0.26 dB	−8.31%
Demichelis Cut, frame 0 (DC)	1.05 dB	−29.25%	0.26 dB	−8.06%	**0.30 dB**	**−9.14%**	0.29 dB	−8.83%
Laura (LAURA)	2.26 dB	−30.35%	0.15 dB	−2.76%	0.27 dB	−4.78%	**0.32 dB**	**−5.62%**
Seagull (SEAGULL)	2.81 dB	−42.78%	0.22 dB	−4.89%	0.31 dB	−6.82%	**0.43 dB**	**−9.21%**
Bikes (BIKES)	0.81 dB	−18.50%	0.11 dB	−2.91%	**0.13 dB**	**−3.34%**	0.11 dB	−2.97%
Danger de Mort (DANGER)	0.61 dB	−14.67%	0.10 dB	−2.72%	**0.11 dB**	**−2.94%**	0.10 dB	−2.75%
Flowers (FLOWERS)	0.17 dB	−4.10%	0.03 dB	−0.75%	**0.03 dB**	**−0.77%**	0.03 dB	−0.67%
Stone Pillars (STONE)	0.25 dB	−6.31%	**0.02 dB**	**−0.61%**	0.02 dB	−0.57%	0.02 dB	−0.57%
Vespa (VESPA)	0.89 dB	−28.73%	0.10 dB	−3.90%	**0.14 dB**	**−5.30%**	0.12 dB	−4.89%
Ankylosaurus & Diplodocus (ANKY)	1.43 dB	−45.35%	0.05 dB	−1.26%	**0.06 dB**	**−4.46%**	0.05 dB	−2.91%
Desktop (DESKTOP)	0.48 dB	−13.73%	0.26 dB	−8.05%	0.25 dB	−7.78%	**0.29 dB**	**−8.93%**
Magnets (MAGNTES)	0.66 dB	−22.95%	0.06 dB	−3.52%	**0.06 dB**	**−5.19%**	0.05 dB	−2.62%
Fountain & Vincent (FOUNTAIN)	1.49 dB	−30.72%	0.17 dB	−4.39%	**0.20 dB**	**−5.31%**	0.20 dB	−5.28%
Friends (FRIENDS)	0.22 dB	−8.10%	0.05 dB	−2.01%	0.06 dB	−2.21%	**0.06 dB**	**−2.36%**
Color chart (COLOR)	1.49 dB	−42.84%	0.27 dB	−11.19%	0.31 dB	−12.30%	**0.32 dB**	**−12.62%**
ISO Chart (ISO)	1.42 dB	−41.35%	0.24 dB	−9.28%	0.27 dB	−10.42%	**0.27 dB**	**−10.47%**
AVG. FOC	1.59 dB	−27.91%	0.29 dB	−6.08%	0.36 dB	−7.47%	**0.37 dB**	**−7.75%**
AVG. UNF	0.83 dB	−23.11%	0.12 dB	−4.22%	**0.14** dB	**−5.06%**	0.14 dB	−4.75%
AVG. ALL	1.08 dB	−24.71%	0.18 dB	−4.84%	**0.21** dB	**−5.86%**	0.21 dB	−5.75%

in terms of average PSNR (BD-PSNR) and bitrate savings. However, comparing the average performance of each method regarding the type of camera models (AVG. FOC and AVG. UNF) it is possible to observe that HEVC-HOP-P is slightly more efficient for the unfocussed model images and HEVC-HOP-B is slightly more efficient for the focused model images. In the case of HEVC-HOP-A only six DoF are available because only three points of correspondence are transmitted. When compared HEVC-HOP-A to HEVC-HOP-P, the bitrate savings gains relatively to HEVC-SS are reduced to 4.84% (28.12% relatively to HEVC) on average considering all tested LF images, which may be due to the fact that HEVC-HOP-A is not able to compensate for perspective changes. However, in terms of computational complexity HEVC-HOP-A is approximately 4.5 times faster than HEVC-HOP-P. Note that none of the implementations is optimized in terms of computational complexity; therefore, the reported values for comparison may vary. It is worthwhile mentioning that for some cases (e.g., STONE test image) HEVC-HOP-A can outperform both eight DoF GTs. In HEVC-HOP-P four correspondence points are always encoded, even if only three are necessary. Since in some cases, more information might be transmitted to describe the same GT, HEVC-HOP-P is, for this particular test image, less efficient than HEVC-HOP-A.

Regarding the computational complexity, a study was performed using the image VESPA, from the EPFL dataset of LF images. This image was encoded and decoded using the codecs, HEVC, HEVC-SS, HEVC-HOP-A, HEVC-HOP-P and HEVC-HOP-B, with QP $= 32$. These tests were performed using a PC equipped with an Intel Xeon CPU E3-1240 Vand 24GB of RAM, running Ubuntu 16.04. The obtained running time to encode and decode each image is depicted in Table IV. The computational complexity of the proposed schemes must be compared to HEVC-SS, as it is used as reference. The HEVC-SS complexity is equivalent to encoding a P-Slice in HEVC [30]. As can be seen from Table IV, the proposed algorithm increases the computational burden at the encoder side, where HEVC-HOP-A, HEVC-HOP-P and HEVC-HOP-B are 1.51, 6.84 and 5.23 times more complex than HEVC-SS, respectively. However, at the decoder the running time is reduced in relation to HEVC-SS. As the proposed method uses a more efficient prediction, the LF image encoder creates a lower number of partitions than the HEVC-SS.

One of the most important advantages of the proposed prediction method is the ability to choose between the LOP stage and the HOP stage for each image block. This decision is taken based on RDO criteria, which allows the proposed HEVC-HOP to outperform HEVC-SS in all cases. In Table II it is possible to observe that despite the HOP stage of the proposed prediction method being used more frequently than the LOP stage, there is always a considerable part of the image that is encoded using only the LOP prediction mode. However, the fact that the HOP stage is used more often than the LOP stage alone indicates that, although the proposed method HEVC-HOP-P requires additional overhead for transmitting the prediction information, i.e., one flag (F_{HOP}) and four vectors ($\{\vec{v}_{HOP_n}\}$), it is very efficient in reducing the distortion between the current block and prediction block, therefore reducing the RD cost. An example of this can be seen in Fig. 6 where a comparison between the generated prediction block using either the LOP method or the proposed HOP method is shown. In this example, the prediction block generated using the

TABLE II
AVERAGE PREDICTION MODE USAGE ACROSS THE FOUR QPS, IN PERCENTAGE OF PIXELS, FOR THE HEVC-HOP-P CASE

Image	DC, Planar and Directional	Proposed prediction method		SS-Skip [30]
		LOP stage	HOP stage	
PT0	**57.76%**	3.32%	22.81%	16.12%
PT150	26.53%	4.39%	**51.27%**	17.62%
DS	27.35%	3.84%	**40.31%**	28.51%
DC	27.25%	1.93%	**44.04%**	26.78%
LAURA	21.51%	10.25%	**45.22%**	23.03%
SEAGULL	13.87%	9.27%	**41.81%**	35.05%
BIKES	**44.12%**	5.69%	33.44%	16.75%
DANGER	**49.28%**	5.99%	31.15%	13.59%
FLOWERS	**81.54%**	2.63%	10.55%	5.28%
STONE	**82.66%**	2.75%	9.60%	4.99%
VESPA	**38.42%**	7.94%	30.03%	23.61%
ANKY	24.39%	10.89%	30.26%	**34.46%**
DESKTOP	**57.50%**	2.67%	15.68%	24.14%
MAGNETS	28.88%	9.57%	28.42%	**33.14%**
FOUNTAIN	30.12%	8.38%	**37.66%**	23.84%
FRIENDS	**78.01%**	3.08%	11.35%	7.57%
COLOR	15.62%	8.94%	**38.14%**	37.30%
ISO	11.40%	11.29%	**41.57%**	35.75%

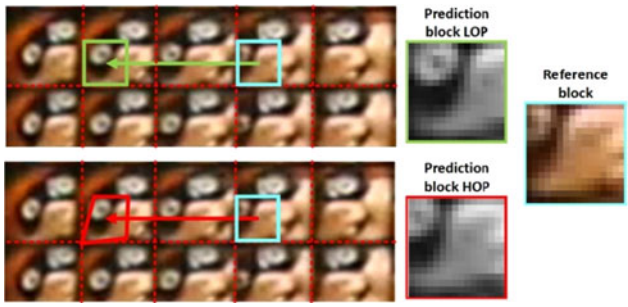


Fig. 6. Comparison between the prediction block generated by LOP and HOP stages.

proposed HOP stage has a lower RD cost ($J_{HOP} = 5080$) than the prediction block using LOP ($J_{LOP} = 7851$), despite the extra bits necessary to convey to the decoder the GT parameters ($\{\vec{v}_{HOP_n}\} = \{(3;0),(4;0),(2;-1),(2;1)\}$).

3) Experimental Results for Rendered SAIs: To further evaluate the compression efficiency of the proposed HEVC-HOP-P, the objective quality of the SAIs extracted from the encoded LF images was tested. The experimental methodology adopted in [23] for the Lytro Illum unfocused camera setup was used. In this case, the PSNR-Y is calculated as an average of the PSNR-Y of 13 × 13 SAIs. This average PSNR-Y compares reference and reconstructed (i.e., encoded and decoded) SAIs. The processing chain designed to generate the SAIs consists in converting the hexagonal lenslet LF image to a square lenslet LF image (with 15x15 pixels per MI) and then, extracting one pixel, in a fixed position from each square MI, to render each SAI. For this case only the images from [23] have been used. The only difference to the methodology in [23] is that, instead of using fixed compression ratios, the results are calculated using the reconstructed images attained with fixed QPs (22, 27, 32 and 37), i.e., the number of bits is the same as in the previously used methodology.

TABLE III
BD-PSNR-Y AND BD-RATE RESULTS COMPARING HEVC, HEVC-SS (2 DOF) AND HEVC-HOP-P (8 DOF) USING THE TESTING METHODOLOGY OF [23]

Image	HEVC-SS vs HEVC		HEVC-HOP-P vs HEVC-SS	
	BD-PSNR-Y	BD-RATE	BD-PSNR-Y	BD-RATE
BIKES	0.78 dB	−20.83%	0.08 dB	−2.41%
DANGER	0.57 dB	−16.08%	0.08 dB	−2.41%
FLOWERS	0.17 dB	−4.67%	0.02 dB	−0.69%
STONE	0.27 dB	−8.43%	−0.03 dB	0.95%
VESPA	0.66 dB	−32.79%	0.09 dB	−5.34%
ANKY	1.48 dB	−62.57%	0.11 dB	−6.87%
DESKTOP	0.33 dB	−15.23%	0.10 dB	−5.19%
MAGNETS	0.64 dB	−37.90%	0.08 dB	−5.97%
FOUNTAIN	1.22 dB	−35.38%	0.12 dB	−4.34%
FRIENDS	0.15 dB	−10.40%	0.03 dB	−1.72%
COLOR	1.26 dB	−58.40%	0.18 dB	−12.86%
ISO	1.49 dB	−45.39%	0.26 dB	−10.81%
AVG.	0.75 dB	−29.01%	0.09 dB	−4.81%

TABLE IV
CODEC COMPUTATIONAL COMPLEXITY COMPARISON

Encoder	HEVC	HEVC-SS	HEVC-HOP-A	HEVC-HOP-P	HEVC-HOP-B
Run time (h)	0.06	3.88	5.88	26.54	20.30
vs HEVC-SS	0.02	1	1.51	6.84	5.23
Decoder	HEVC	HEVC-SS	HEVC-HOP-A	HEVC-HOP-P	HEVC-HOP-B
Run time (s)	1.70	33.56	30.54	29.98	28.56
vs HEVC-SS	0.05	1	0.91	0.89	0.85

From the results presented in Table III it is possible to observe that there is a coherence between the results for the encoded LF images and the rendered SAIs. The average bitrate savings achieved by HEVC-HOP-P relative to HEVC-SS are very similar for both cases. Nevertheless, as the RD cost was not optimized on each SAI, the results are not exactly the same. As previously explained, when using the HEVC-HOP-P method the LF image is encoded "as is", without the need to know any information about the used lenslet based LF camera. This may explain the bitrate increase for image STONE, as the proposed method calculates the RD cost based on the lenslet LF image, instead of the generated view or SAI.

Additionally, a comparison between the proposed HEVC-HOP-P and the state-of-the-art method [24] was performed. In [24], a comparison in relation to JPEG, using the EPFL dataset, reports a gain of 4.54 dB in the BD-PSNR. Similarly, this image set were encoded by the proposed HEVC-HOP-P, with QPs {22, 27, 32, 37}, achieving a BD-PSNR gain of 4.83 dB, in relation to JPEG. For a wider QP range {17, 22, 27, 32, 37}, the gain of HEVC-HOP-P decreases to 4.34 dB. Thus, it is fair to assume that the proposed HEVC-HOP-P has a very similar performance to the state-of-the-art method [24].

4) Results for Different Lenslet LF Camera Models: In general, the bitrate savings across the different codecs when compared to HEVC are higher when encoding LF images captured with cameras using a focused LF camera model. This is possible to see when comparing the average bitrate savings for the

LF images captured with an unfocused camera model or the focused camera model in Table II. This occurs because HEVC-SS and HEVC-HOP are based on matching prediction tools. In the focused images, the MIs are focused, therefore sharper than the unfocussed images. In sharper MI, more prominent features exist and therefore the block matching is more reliable [34]. Additionally, since the incident light in the camera's sensor in the unfocused case is focused at infinity, the disparity between MIs tends to be zero, which means that theoretically no perspective compensation can be matched. This can be justified by the noticeable lower relative prediction mode usage, shown in Table II, for the proposed prediction method as well as SS-Skip for most LF images captured with unfocused camera models.

The proposed HOP model is also more suited to adapt to non-rectangular shape MIs, e.g., hexagonal and circular shape, when compared to LOP model based methods. This happens because the corners of the prediction blocks, when using the proposed HOP model, are flexible to adapt for different block shapes.

VI. Conclusion

In this paper, a HOP mode for LF image coding was proposed, using geometric transformations of up to eight DoF. The proposed HOP mode is a two-stage block-wise approach that is able to achieve RD efficiency gains, relative to a LOP state-of-the-art solution for LF image coding and HEVC. These gains occur, regardless of the LF camera model, MI and LF image resolution and microlens array type. Experimental results show average bitrate savings of 5.86% and 28.81%, when compared to a LOP state-of-the-art solution and HEVC, respectively, across different types of LF images, when using the Projective GT. It is also possible to conclude that, the GTs with eight DoF, namely Projective and Bilinear, are generally more efficient than Affine GT in a RD sense.

An additional testing methodology, based on the SAI objective quality, was also used to confirm the bitrate savings when comparing HOP with LOP tools. In this case the average bitrate savings achieved is 4.81% and 31.77% when comparing the HOP mode with the state-of-the-art LOP solution and HEVC, respectively.

Future work will include the investigation of GT parameter prediction techniques and optimal HOP model selection, aiming to combine them in the same codec. Additionally, entropy encoding improvements will also be considered, namely, the binarization of the HOP model vectors.

References

[1] T. Georgiev and A. Lumsdaine, "Rich image capture with plenoptic cameras," in *Proc. IEEE Int. Conf. Comput. Photography*, Cluj-Napoca, Romania, 2010, pp. 1–8.

[2] J. Arai, "Integral three-dimensional television (FTV Seminar)," ISO/IEC JTC1/SC29/WG11 MPEG2014/N14552, Sapporo, Japan, Jul. 2014.

[3] X. Xiao, B. Javidi, M. Martinez-Corral, and A. Stern, "Advances in Proc. three-dimensional integral imaging: sensing, display, and applications," *Appl. Opt.*, vol. 52, no. 4, pp. 546–560, Feb. 2013.

[4] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *Proc. IEEE Int. Conf. Comput. Photography*, 2009, pp. 1–8.

[5] C. Hahne, A. Aggoun, S. Haxha, V. Velisavljevic, and J. C. J. Fernández, "Light field geometry of a standard plenoptic camera," *Opt. Express*, vol. 22, no. 22, pp. 26659–26673, Nov. 2014.

[6] "JPEG PLENO abstract and executive summary," ISO/IEC JTC 1/SC 29/WG1 N6922, Sydney, NSW, Australia, Feb. 2015.

[7] S. Kumar and R. C. Jain, "Low complexity fractal-based image compression technique," *IEEE Trans. Consum. Electron.*, vol. 43, no. 4, pp. 987–993, Nov. 1997.

[8] H. Huang, J. W. Woods, Y. Zhao, and H. Bai, "Control-point representation and differential coding affine-motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 10, pp. 1651–1660, Oct. 2013.

[9] T. Wiegand, E. Steinbach, and B. Girod, "Affine multipicture motion-compensated prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 2, pp. 197–209, Feb. 2005.

[10] M. Tok, M. Esche, A. Glantz, A. Krutz, and T. Sikora, "A parametric merge candidate for high efficiency video coding," in *Proc. Data Compression Conf.*, 2013, pp. 33–42

[11] A. Krutz, A. Glantz, M. Tok, M. Esche, and T. Sikora, "Adaptive global motion temporal filtering for high efficiency video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1802–1812, Dec. 2012.

[12] M. Tok, A. Glantz, A. Krutz, and T. Sikora, "Feature-based global motion estimation using the Helmholtz principle," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 1561–1564.

[13] R. C. Kordasiewicz, M. D. Gallant, and S. Shirani, "Affine motion prediction based on translational motion vectors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 10, pp. 1388–1394, Oct. 2007.

[14] X. San, H. Cai, J. G. Lou, and J. Li, "Multiview image coding based on geometric prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1536–1548, Nov. 2007.

[15] K. Müller *et al.*, "3D high-efficiency video coding for multi-view video and depth data," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3366–3378, Sep. 2013.

[16] R. J. S. Monteiro, N. M. M. Rodrigues, and S. M. M. Faria, "Disparity compensation using geometric transforms," in *Proc. 3DTV-Conf., True Vis. Capture, Transmiss. Display 3D Video*, 2014, pp. 1–4.

[17] R. J. S. Monteiro, N. M. M. Rodrigues, and S. M. M. Faria, "Geometric transforms and reference picture list optimization for efficient disparity compensation," in *Proc. Int. Conf. Image Process. Theory, Tools Appl.*, 2015, pp. 370–375.

[18] A. Aggoun, "A 3D DCT compression algorithm for omnidirectional integral images," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, vol. 2, pp. II–II.

[19] M. C. Forman and A. Aggoun, "Quantisation strategies for 3D-DCT-based compression of full parallax 3D images," in *Proc. Int. Conf. Image Process. Its Appl.*, 1997, vol. 1, pp. 32–35.

[20] A. Aggoun, "Compression of 3D integral images using 3D wavelet transform," *J. Display Technol.*, vol. 7, no. 11, pp. 586–592, Nov. 2011.

[21] F. Dai, J. Zhang, Y. Ma, and Y. Zhang, "Lenselet image compression scheme based on subaperture images streaming," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 4733–4737.

[22] A. Vieira, H. Duarte, C. Perra, L. Tavora, and P. Assuncao, "Data formats for high efficiency coding of Lytro-Illum light fields," in *Proc. Int. Conf. Image Process. Theory, Tools Appl.*, 2015, pp. 494–497.

[23] "ICME 2016 grand challenge: Light-field image compression." [Online]. Available: http://mmspg.epfl.ch/files/content/sites/mmspl/files/shared/LF-GC/CFP.pdf, Accessed on: Dec. 1, 2016.

[24] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2016, pp. 1–4.

[25] J. Boyce, J. Chen, E. Alshina, G. J. Sullivan, and J.-R. Ohm, "Algorithm description of Joint Exploration Test Model 3 (JEM3)," ISO/IEC JTC1/SC29/WG11/N16276, Jun. 2016.

[26] M. Magnor and B. Girod, "Data compression for light-field rendering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 3, pp. 338–343, Apr. 2000.

[27] B. Girod, C.-L. Chang, P. Ramanathan, and X. Zhu, "Light field compression using disparity-compensated lifting," in *Proc. Int. Conf. Multimedia Expo*, 2003, vol. 1, pp. 373–376.

[28] Y. Li, M. Sjöström, and R. Olsson, "Coding of plenoptic images by using a sparse set and disparities," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2015, pp. 1–6.

[29] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Scalable coding of plenoptic images by using a sparse set and disparities," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 80–91, Jan. 2016.

[30] C. Conti, L. D. Soares, and P. Nunes, "HEVC-based 3D holoscopic video coding using self-similarity compensated prediction," *Signal Process., Image Commun.*, vol. 42, pp. 59–78, Mar. 2016.

[31] C. Conti, P. Nunes, and L. D. Soares, "HEVC-based light field image coding with bi-predicted self-similarity compensation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2016, pp. 1–4.
[32] L. F. R. Lucas *et al.*, "Locally linear embedding-based prediction for 3D holoscopic image coding using HEVC," in *Proc. Eur. Signal Process. Conf.*, Lisbon, Portugal, 2014, pp. 11–15.
[33] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Coding of focused plenoptic contents by displacement intra prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 7, pp. 1308–1319, Jul. 2016.
[34] Y. Li, R. Olsson, and M. Sjöström, "Compression of unfocused plenoptic images using a displacement intra prediction," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2016, pp. 1–4.
[35] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[36] P. S. Heckbert, "Fundamentals of texture mapping and image warping," Masters' thesis, Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley, CA, USA, Rep. UCB/CSD-89-516, Jun. 1989.
[37] J. Jain and A. Jain, "Displacement measurement and its application in interframe image coding," *IEEE Trans. Commun.*, vol. 29, no. 12, pp. 1799–1808, Dec. 1981.
[38] C.-H. Cheung and L.-M. Po, "A novel cross-diamond search algorithm for fast block motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 12, pp. 1168–1177, Dec. 2002.
[39] J. R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards—including high efficiency video coding (HEVC)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1669–1684, Dec. 2012.
[40] F. Bossen, B. Bross, K. Suhring, and D. Flynn, "HEVC complexity and implementation analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1685–1696, Dec. 2012.
[41] A. Agooun *et al.*, "Acquisition, processing and coding of 3D holoscopic content for immersive video systems," in *Proc. 3DTV-Conf., True Vis.-Capture, Transmiss. Dispaly 3D Video*, Aberdeen U.K., 2013, pp. 1–4.
[42] G. Todor, "Gallery and lightfield data." [Online]. Available: http://www.tgeorgiev.net/Gallery/, Accessed on: Dec. 1, 2016.



Ricardo J. S. Monteiro (M'14) received the Graduate degree in electrical engineering in 2012 and the M.Sc. degree from the Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Leiria, Leiria, Portugal, in 2014. He is currently working toward the Ph.D. degree at the University Institute of Lisbon, Lisbon, Portugal.

Since 2010, he has been with the Instituto de Telecomunicações, Portugal, as a Researcher. His current research interests include image and video processing, namely, multiview video and light field image processing and coding.



Paulo J. L. Nunes (M'98) received the Graduate, M.Sc., and Ph.D. degrees in electrical and computers engineering from Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisbon, Portugal, in 1992, 1996, and 2007, respectively.

He is currently an Assistant Professor in the Information Science and Technology Department, University Institute of Lisbon, Portugal, and a member of the Research Staff of Instituto de Telecomunicações, Portugal. His current research interests include 2-D/3-D image and video processing and coding, namely, light field image and video processing and coding.

Dr. Nunes has coordinated and participated in various national and international (EU) funded projects and has acted as a Project Evaluator for the European Commission. He acts often as a Reviewer for various conferences and journals and a member of the program and organizing committees of various international conferences. He has contributed more than 60 papers.



Nuno M. M. Rodrigues (M'99) received the Graduate degree in electrical engineering in 1997, the M.Sc. degree from the Universidade de Coimbra, Coimbra, Portugal, in 2000, and the Ph.D. degree from the Universidade de Coimbra, in 2009, in collaboration with the Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.

Since 1997, he has been with the Department of Electrical Engineering, Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Leiria, Leiria, Portugal. Since 1998, he has been with the Instituto de Telecomunicações, Portugal, where he is currently a Senior Researcher. He has coordinated and participated as a Researcher in various national and international funded projects. His current research interests include digital signal and image processing, namely, image and video compression, point cloud and light field image and video compression, medical image compression, and many-core programming.



Sérgio M. M. Faria (M'93–SM'08) was born in Portugal in 1965. He received the Engineering degree and the M.Sc. degree, both in electrical engineering, from Universidade de Coimbra, Portugal, in 1988 and 1992, respectively, and the Ph.D. degree in electronics and telecommunications from the University of Essex, Colchester, U.K., in 1996.

Since 1990, he has been a Professor with the Department of Electrical Engineering, Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Leiria, Portugal. He has collaborated in master courses with the Faculty of Science and Technology and with the Faculty of Economy of Universidade de Coimbra, Portugal. He is an Auditor with A3ES organization for electrical and electronic engineering courses in Portugal. He is a Senior Researcher with Instituto de Telecomunicações.

Dr. Faria's research interests include 2-D/3-D image and video processing and coding, motion representation, and medical imaging. In this field, he has published 1 book, edited 2 books, and authored 7 book chapters, 23 journal papers, and 117 referred conference papers. He has been participating and is responsible for several, national and international (EU), funded projects.

He is an Area Editor of *Signal Processing: Image Communication*. He has been a Scientific and Program Committee Member of many international conferences. He is a Reviewer for several international scientific journals and conferences (IEEE, IET, and EURASIP). In 2014, he received the title of "Agregado" by Instituto Superior Técnico, University of Lisbon.

Light Field Compression With Homography-Based Low-Rank Approximation

Xiaoran Jiang, Mikaël Le Pendu, Reuben A. Farrugia, and Christine Guillemot, *Fellow, IEEE*

***Abstract*—This paper describes a light field compression scheme based on a novel homography-based low-rank approximation method called HLRA. The HLRA method jointly searches for the set of homographies best aligning the light field views and for the low-rank approximation matrices. The light field views are aligned using either one global homography or multiple homographies depending on how much the disparity across views varies from one depth plane to the other. The light field low-rank representation is then compressed using high efficiency video coding (HEVC). The best pair of rank and quantization parameters of the coding scheme, for a given target bit rate, is predicted with a model defined as a function of light field disparity and texture features. The results are compared with those obtained by directly applying HEVC on the light field views restructured as a pseudovideo sequence. The experiments using different datasets show substantial peak signal to noise ratio (PSNR)-rate gain of our compression algorithm, as well as the accuracy of the proposed parameter prediction model, especially for real light fields. A scalable extension of the coding scheme is finally proposed.**

***Index Terms*—Compression, homography, light fields, low rank approximation.**

I. Introduction

LIGHT field (LF) imaging has emerged as a very promising technology in the field of computational photography. Many acquisition devices have been recently designed to capture light fields, going from arrays of cameras capturing the scene from slightly different viewpoints [1], to single cameras mounted on moving gantries and plenoptic cameras. Plenoptic cameras are becoming commercially available using arrays of micro-lenses placed in front of the photosensor to obtain angular information about the captured scene [2], [3]. Compared to classical 2D imaging, light fields capture the intensity values of light rays interacting with the scene. The recorded flow of rays is in the form of large volumes of data retaining both spatial and angular information of a scene which enables a variety of post-capture processing, such as re-focusing, extended focus, different viewpoint rendering and depth estimation, from a single exposure [1], [2], [4]. For a comprehensive overview of light field image processing techniques, please refer to [5].

Given the very large volume of high-dimensional data, the design of efficient compression schemes of light fields is a key challenge for practical use of this technology. First methods for compressing synthetic light fields appeared in the late 90's essentially based on classical coding tools as vector quantization followed by Lempel-Ziv (LZ) entropy coding [4] or wavelet coding as in [6] and [7], yielding however limited compression performances (compression factors not exceeding 20 for an acceptable quality). Predictive schemes inspired from video compression methods have also been naturally investigated, adding specific prediction modes in schemes inspired from H.264 and MVC, as in [8] and [9]. Motivated by the objective of random access and progressive decoding, which is not enabled by predictive schemes, the authors in [10] describe another approach using a wavelet transform applied in the 4 dimensions of the light field. The scheme naturally inherits the scalable and progressive properties of wavelet-based coding schemes. The authors in [11] instead use a Principal Component Analysis (PCA) to enable random access to pixels and to support scalability.

The compression of real light fields has recently gained attention thanks to the emergence of capturing devices (plenoptic cameras or rigs of cameras). In this paper, we focus on the compression of real light fields captured by plenoptic cameras. Prior work in this area has followed two main directions: either coding the array of sub-aperture images extracted from the lenslet image as in [12]–[14], or directly encoding the lenslet images captured by plenoptic cameras [15]–[20] with an extension of high efficiency video coding (HEVC) with dedicated prediction modes.

Instead of directly encoding the light field (the array of sub-aperture images or the lenslet image for light fields captured by plenoptic cameras), the authors in [21] consider the focus stack as an intermediate representation of reduced dimension of the light field and encode the focus stack with a wavelet-based scheme. The light field is then reconstructed from the focus stack using the linear view synthesis approach described in [22].

In this paper, we describe a light field compression algorithm based on a low rank approximation exploiting scene and data geometry. In particular, light fields with a dense angular sampling such as those captured by plenoptic cameras are addressed. We consider the coding of the sub-aperture images (i.e., views) already extracted from a lenslet image. Thanks to the high correlation between the views in such light fields, the matrix

Manuscript received November 29, 2016; revised May 5, 2017 and August 16, 2017; accepted August 24, 2017. Date of publication August 30, 2017; date of current version October 23, 2017. This work was supported by the EU H2020 Research and Innovation Programme under Grant Agreement 694122 (ERC Advanced Grant CLIM). The guest editor coordinating the review of this paper and approving it for publication was Dr. Lu Fang. *(Corresponding author: Christine Guillemot.)*

X. Jiang, M. Le Pendu, and C. Guillemot are with the Inria Centre de Recherche Rennes Bretagne Atlantique, Rennes 35042, France (e-mail: xiaoran.jiang@inria.fr; mikael.le-pendu@inria.fr; christine.guillemot@inria.fr).

R. A. Farrugia is with the Department of Communications and Computer Engineering, University of Malta, Msida MSD2080, Malta (e-mail: reuben.farrugia@um.edu.mt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2747078

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

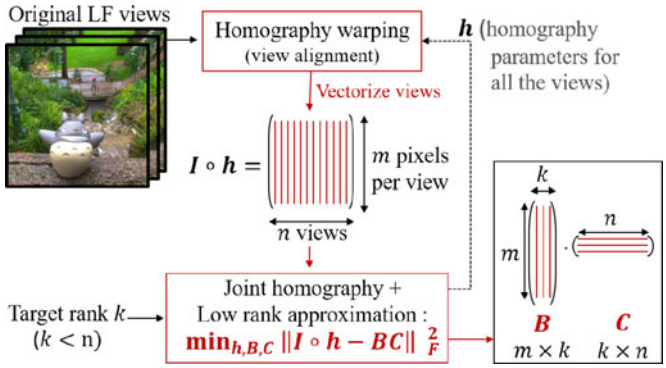


Fig. 1. Overview diagram of the HLRA method.

whose columns are formed by vectorizing each view can be well approximated by a low rank matrix. In addition, a prior alignment of the views with homography warpings increases the correlation and thus improves the low rank approximation. In the proposed method, homography projections are searched for each view in order to obtain the best low rank matrix approximation for a given target rank k (where k is less than the number of views). This joint homography alignment and low rank optimization procedure is illustrated in Fig. 1. In the cases where the scene contains several layers of depth, the method has also been extended to search for one homography for each depth plane. To cope with artifacts at the frontier of two depth planes when performing the homography warpings, a blending of homographies is performed.

The obtained rank k matrix is expressed as a product of a matrix B, containing k basis vectors, with a matrix C containing weighting coefficients as shown in Fig. 1. This decomposition enables an efficient encoding where the k images formed by reshaping each column of B are compressed with HEVC-Intra. The weights contained in C and the homography parameters are also necessary to reconstruct the light field, hence are transmitted using classical entropy coding. Note that this low rank approximation method is based on our earlier work described in [14].

The peak signal to noise ratio (PSNR)-rate performance of the compression scheme depends on two key parameters: the approximation rank and the quantization parameter (QP) of the HEVC encoder. Using a set of training light fields, we learn a model to predict the pair of rank and QP parameters yielding the best PSNR for a given target bit-rate. The model is learned as a function of several input light field features: disparity indicators defined as a function of the decay rate of the SVD values of the original and registered view matrices, as well as texture indicators defined in terms of the decay rate of SVD values computed on the central view. The parameter prediction problem is cast as a multi-output classification problem solved using a Decision Tree ensemble method, namely the Random Forest method [23].

A scalable extension is finally described in which the homography-based low rank model is used to compress the base layer. The reconstructed base layer is used to predict the light field, and the residue for each light field view is encoded by using the proposed low rank based scheme, with or without view alignment.

Our experiments show the advantage of using the proposed joint alignment and low rank optimization rather than first aligning the views independently of the targeted rank. Furthermore, the compression performances of the proposed scheme are assessed against those obtained with two methods applying HEVC inter-coding on pseudo-sequences of sub-aperture images [12], [13]. In the sequel, the method in [12] will be referred to as HEVC-lozenge, while the method in [13] will be called HEVC-pseudo. Thanks to the robustness of our method to noise and color variations across views, substantial PSNR-rate gains are obtained for two different datasets of real light fields: the INRIA dataset [24] and that of the ICME 2016 Grand Challenge (Light-Field Image Compression) [25].

II. Light Fields: Background and Notations

The light rays emitted by a scene and received by an observer at a particular point (x, y, z) in space, at a given instant t, can be described by the 7D plenoptic function $L(x, y, z, \theta, \phi, t, \lambda)$ where (θ, ϕ) are angles giving the orientation of the light rays and λ their wavelength. For a static light field, the $7D$ plenoptic function can be simplified into a 4D representation called 4D light field in [4] and lumigraph in [26], describing the radiance along rays by a function $L(x, y, u, v)$ of 4 parameters at the intersection of the light rays with 2 parallel planes. This simplification is done assuming constant radiance of a light ray from point to point, and given that an R, G, B sampling of the wavelength is performed by the color filters coupled with the CCD sensors.

The light field can be seen as capturing an array of viewpoints (called sub-aperture images in particular in the case of microlens based capturing devices) of the imaged scene with varying angular coordinates u and v. The different views will be denoted here $I_{u,v} \in \mathbb{R}^{X \times Y}$, where X and Y represent the vertical and horizontal dimensions of each sub-aperture image. Each sub-aperture image corresponds to a fixed pair of (u, v) coordinates. In the following, the notation $I_{u,v}$ for the different views (or sub-aperture images) is simplified as I_i with a bijection between (u, v) and i. The complete light field can hence be represented by the matrix $I \in \mathbb{R}^{m \times n}$:

$$I = [\text{vec}(I_1) \mid \text{vec}(I_2) \mid ... \mid \text{vec}(I_n)], \quad (1)$$

with $\text{vec}(I_i)$ being the vectorized version of the sub-aperture image I_i, and where m is the number of pixels in each view ($m = X \times Y$) and n is the number of views in the light field.

III. Related Work

A. Low-Rank Approximation

Many algorithms taking advantage of the low rank property of a matrix have been developed in the recent literature. For instance, the methods in [27]–[34] tackle the problem of completing a low rank matrix from a subset of its entries. A closely related problem referred to as robust principal component analysis (RPCA) consists in decomposing a matrix as the sum of a

low rank and a sparse matrix. The authors in [35] solved this problem using an augmented Lagrangian multiplier method. Note that although those problems are NP-hard in general and involve non-convex optimization (the rank of a matrix being non-convex), recent theoretical papers [36]–[38] have shown that under surprisingly broad conditions, a convex relaxation replacing the rank by the nuclear norm leads exactly to the same unique solution.

Inspired by the RPCA, a low rank approximation model has been considered in the RASL method [39] for aligning correlated images. In their model, each input image is warped by a homography projection. The homography parameters are determined in order to optimize the low rank and sparse decomposition of the matrix formed by concatenating the vectorized warped images. In the context of light filed compression, however, this sparse and low rank decomposition approach does not necessarily yield a compact representation of the original light field. Although useful for ignoring inconsistent pixels between views in the RASL alignment, the sparse term still contains important visual features of the light field (e.g., specular light, disocclusion) which essentially consist of high frequencies. Therefore, encoding this term is likely to degrade the overall coding performance. The authors in [40] have nevertheless shown the efficiency of such a compression scheme for encoding videos captured with a fixed camera. In that case the sparse term only captures the moving objects which can be efficiently compressed with a classical video encoder. The static background is essentially contained in the low rank term with a typical rank of 1 or close to 1, and can be encoded at a very low cost.

Note also that for encoding efficiently a low rank matrix A of size $m \times n$ and of rank $k \ll \min(m,n)$, a preliminary factorization step of the form $A = BC$ (where $B \in \mathbb{R}^{m \times k}$ and $C \in \mathbb{R}^{k \times n}$) must be performed. Such a factorization can be obtained with a singular value decomposition. Alternatively, in the SLRMA compression method [41], a similar factorization is found by solving an optimization problem which additionally constrains the sparsity of the matrix B in a given dictionary.

B. Light Fields Compression

In this section, we focus on prior work dealing with the compression of light fields captured by plenoptic cameras. The methods proposed in the literature can be classified into two categories: those which aim at directly compressing the raw lenslet data after de-vignetting and demosaicing (e.g., [15]–[20]) and those which compress the sub-aperture images extracted from the lenslet data (e.g., [12], [13]).

Most solutions proposed for directly encoding the lenslet data aim at exploiting spatial redundancy or self-similarity between the micro-images. The micro-image is the set of pixels behind each micro-lens, and is also sometimes called elemental image. Spatial prediction modes have thus been proposed for unfocused cameras in [15] and [42] based on a concept of self-similarity compensated prediction or using locally linear embedding techniques in [43]. These self-similarity prediction modes have been further extended to bi-directional prediction in [18] and [19]. The authors in [16] introduce a bi-directional spatial prediction mode in HEVC for encoding elemental images captured by a focused 2.0 camera [44] which has been further extended in [20]) for unfocused 1.0 cameras. While the elemental images (EI) in the focused 2.0 plenoptic cameras can be seen as a cropped multi-view image from one viewing angle, the EI produced by the unfocused 1.0 cameras capture angular information of one point in space. A scalable extension of HEVC-based scheme is also proposed in [45] where a sparse set of micro-lens images (also called elemental images) is encoded in a base layer. The other elemental images are reconstructed at the decoder using disparity-based interpolation and inpainting. The reconstructed images are then used to predict the entire lenslet image and a prediction residue is transmitted yielding a multi-layer scheme. The authors in [17] instead partition the raw light field data into tiles which are then encoded as a pseudo-video sequence using HEVC.

A second category of method consists in encoding the set of sub-aperture images (or views) which can be extracted from the lenslet images after de-vignetting, demosaicing and alignment of the micro-lens array on the sensor, following the raw data decoding pipeline described in [46]. The author of [47] exploits inter-view correlation by using homography and 2D warping to predict views. Homographies are computed via Random Sample Consensus (RANSAC) [48]. The authors in [12] form a pseudo-sequence by using a lozenge scanning order and encode this pseudo-sequence using HEVC inter-coding. In [13], a coding order and a prediction structure inspired from those used in the multi-view coding (MVC) coding standard is proposed, showing significant performance gains compared with HEVC-Intra.

Here, we consider instead a very different approach which aims at reducing the dimensionality of the data using low rank approximations prior encoding.

IV. Homography-Based Low Rank Approximation

The error introduced by the low rank approximation model depends on how well the sub-aperture images are aligned. We hence propose to search for the homographies minimizing the low rank approximation error for a targeted rank.

Let I_i and I_j be two sub-aperture images for which we assume there exists an invertible homography transformation h_i, such that

$$I_j(x,y) = (I_i \circ h_i)(x,y) = I_i(h_i(x,y)). \tag{2}$$

A homography transformation h_i can be characterized by a 3×3 matrix H_i which transforms each coordinates (x,y) in I_i into the coordinates $(\frac{x_H}{w_H}, \frac{y_H}{w_H})$, where

$$[x_H, y_H, w_H]^\top = H_i \cdot [x, y, 1]^\top. \tag{3}$$

However, without loss of generality, the last element $H_{i(3,3)}$ can be fixed to 1. The eight remaining elements are then sufficient to parametrize the homography.

Let h be the set of homographies associated to each view of the light field. In what follows, we will consider h as the matrix $[h_1 \mid ... \mid h_n]$ where $h_1, ..., h_n$ are vectors of size 8×1 whose elements are the homography parameters. The low rank optimization problem is then formulated as

$$\text{argmin}_{h,B,C} \left\| I \circ h - BC \right\|_F^2, \tag{4}$$

where $\|.\|_F$ is the Frobenius norm, $B \in \mathbb{R}^{m\times k}, C \in \mathbb{R}^{k\times n}$ $(k < n)$, and $I \circ h$ stands for the matrix containing all views aligned using homographies $h_1, ...h_n$ and can be written as

$$I \circ h = [\text{vec}(I_1 \circ h_1) \mid ... \mid \text{vec}(I_n \circ h_n)]. \quad (5)$$

Note that for aligning all the views, only $n-1$ homographies would be sufficient assuming that one view (e.g., the central view) is not warped and is used as reference for aligning the other views. However, this would require constraining the minimization problem (4) to ensure that the homography of the central view is equal to the identity. For simplicity, we did not consider such a constraint in our formulation.

A. Linear Approximation

Minimizing (4) is non trivial due to the non linearity of the term $I \circ h$. Nevertheless, when the change in h is small, we can approximate it by local linearity as follows:

$$I \circ (h + \Delta h) \approx I \circ h + \sum_{i=1}^{n} J_i \Delta h_i \epsilon_i^\top, \quad (6)$$

where $\Delta h = [\Delta h_1 \mid ... \mid \Delta h_n]$, J_i is the Jacobian matrix of the warped and vectorized sub-aperture image, $\text{vec}(I_i \circ h_i)$, with respect to the parameters of h_i (i.e., $J_i = \frac{\partial}{\partial \zeta}\text{vec}(I_i \circ \zeta)|_{\zeta = h_i}$). And ϵ_i is a $n \times 1$ vector with element i equal to 1 and all the other elements equal to 0.

B. Iterative Minimization

The minimization problem in (4) is iteratively solved by updating alternatively the matrices B and C and the homographies $h_1, ..., h_n$. Each homography h_i is first initialized so that the corresponding 3×3 matrix H_i is equal to the identity.

1) Given h fixed, B and C are found by computing the singular value decomposition $I \circ h = U\Sigma V^\top$. Then B is set as the k first columns of $U\Sigma$ and C is set as the k first rows of $V^\top$, so that BC is the closest k-rank approximation of $I \circ h$.
2) h is updated by solving (4) for B and C fixed. By noting the updated homography parameter matrix $h' = h + \Delta h$, the minimization (4) becomes:

$$h' = h + \underset{\Delta h}{\text{argmin}} \|I \circ (h + \Delta h) - BC\|_F^2 \quad (7)$$

Given the approximation in (6), the problem is then to find:

$$\Delta h = \underset{\Delta h}{\text{argmin}} \|I \circ h - BC + \sum_{i=1}^{n} J_i \Delta h_i \epsilon_i^\top\|_F^2 \quad (8)$$

This problem is independently solved for each column Δh_i of Δh and can be equivalently stated as:

$$\forall i,\ \Delta h_i = \underset{\Delta h_i}{\text{argmin}} \|(I \circ h - BC)\epsilon_i + J_i \Delta h_i\|_F^2 \quad (9)$$

This is a linear least squares problem with solution:

$$\forall i,\ \Delta h_i = J_i^\dagger (BC - I \circ h)\epsilon_i \quad (10)$$

where $J_i^\dagger$ denotes the Moore-Penrose pseudoinverse of J_i.

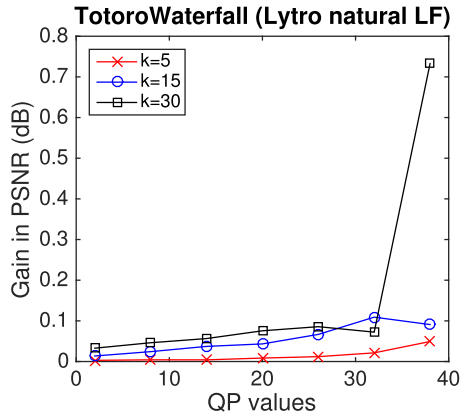


Fig. 2. PSNR gain when the matrix C is re-calculated to account for the compression errors of the matrix B. Alignment has been performed using one homography per view.

C. Recalculating C to Account for Quantization Errors in B

Since the matrix B will need to be compressed to be transmitted to the receiver side, the receiver will obtain a matrix B' with compression artifacts. To reduce the impact of the compression (i.e., quantization) errors on the light field reconstruction, the matrix C is recalculated to account for these quantization errors, as follows:

$$C' = \underset{C}{\text{argmin}} \|I \circ h - B'C\|_F^2 = (B')^\dagger (I \circ h). \quad (11)$$

In practice, this adaptation of C to the compression artifacts of the matrix B can increase the PSNR by about 1 dB when strong compression is applied (e.g., QP = 38). Fig. 2 shows the PSNR gain of this adaptation for the light field "TotoroWaterfall" (*cf.* Fig. 6). The PSNR gains are shown for the case where one homography per view is applied.

V. Low Rank Approximation Using Multiple Homographies

When the disparity varies from one depth plane to another, one global homography per view is not sufficient to well align the whole views. For such light fields, the above homography-based low rank approximation method can be extended to the case where different homographies are computed for different depth planes segmented thanks to a scene depth map.

In order to avoid the high transmission cost of one depth map per view, we instead consider a single depth map of the central view which can be directly estimated using methods as proposed in [49]–[51]. The same depth map is then used for all the views. Note that for the type of light fields addressed, with a dense angular sampling but a limited baseline, using only the depth map of the central view results in a limited loss of accuracy for extracting the depth planes and for computing the corresponding homographies.

We apply one homography to each depth plane p of the sub-aperture image I_i. In what follows, we note h_i^p the vector containing the corresponding homography parameters. Blending is required to naturally mix the depth planes. Here, instead of blending the pixel values, we blend the homographies, which



Fig. 3. Separation on two depth planes: depth map (the left column) and the weights w^1 and w^2 associated to each depth plane (the middle and the right columns).

yields less artifacts at the frontier of depth planes. For that purpose, at each pixel coordinate (x, y), we define the series of weights $\{w^p_{(x,y)}\}_{p\in[\![1,q]\!]}$ that determine the importance of the homography of each depth plane p for this pixel:

$$w^p_{(x,y)} = \begin{cases} 1, & \text{if } D_{(x,y)} \in [\sigma_p + \delta, \sigma_{p+1} - \delta]\,; \\ \frac{D_{(x,y)} - (\sigma_p - \delta)}{2\delta}, & \text{if } |D_{(x,y)} - \sigma_p| < \delta; \\ \frac{(\sigma_{p+1} + \delta) - D_{(x,y)}}{2\delta}, & \text{if } |D_{(x,y)} - \sigma_{p+1}| < \delta; \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

with δ a shallow neighborhood where the blending is applied. Fig. 3 shows the weights w^1 and w^2 associated to each depth plane for the LF "TotoroWaterfall".

The warping (3) is then modified as follows:

$$[x_H, y_H, w_H]^\top = \sum_{p=1}^{q} w^p_{(x,y)} H^p_i \cdot [x, y, 1]^\top . \tag{13}$$

Once the warped images $I \circ h$ are obtained by applying (13) for each sub-aperture image and each depth plane, we compute B and C at each iteration exactly as described in Section IV. Note that we must now determine nq homographies. The size of h is then $8 \times nq$. Similar to (10), each vector of homography parameters h^p_i is updated by adding Δh^p_i computed as:

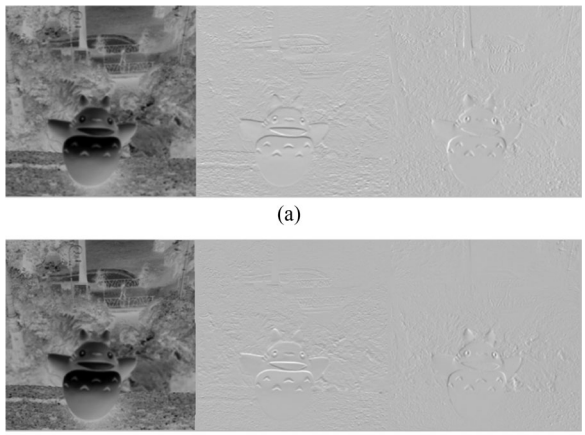
$$\forall i, p,\ \Delta h^p_i = J^{p\,\dagger}_i \left[(BC - I \circ h)\epsilon_i \odot \mathrm{vec}(M^p)\right], \tag{14}$$

with M^p the corresponding binary mask of depth plane p and $\odot$ the Hadamard product.

VI. Compression Algorithm

A. Algorithm Overview

The different steps of the complete compression algorithm are shown in Fig. 5. The low rank representation is compressed by encoding the columns of the matrix B using HEVC Intra coding. However, any encoder could be used to compress the proposed homography-based low rank representation. The columns of the matrix B are first quantized on 16 bits before being encoded using HEVC-Intra coding. The first three columns of the matrix B for the LF "TotoroWaterfall" are shown in Fig. 4. The first column represents low frequency information, whereas the others contain data with high frequency. One can observe that by using homographies to align sub-aperture views, the average image in the first column becomes sharper, and there is less high frequency information remaining in the following columns.



(a)

(b)

Fig. 4. First three columns of the matrix B ("TotoroWaterfall" LF). (a) Without alignment. (b) With alignment, 1 homography per view, $k = 5$.

Besides the matrix B, additional elements need to be transmitted. The coefficients of the matrix C of size $k \times n$, where k and n are the approximation rank and the number of views, are encoded using a scalar quantization on 16 bits and Huffman coding. The $8 \times n \times q$ homography parameters, with q the number of depth planes per view, are encoded the same way. In the case where multiple homograhies are applied, depth-planes need to be segmented. So that the decoder can find the depth planes, as explained in Section V, one depth map is encoded using 8 bit quantization followed by HEVC-Intra (we used QP=32 in the experiments). Percentages of bitrate cost for the different elements are analyzed in Section VII-A4.

B. Model-Based Coding Parameters Prediction

For a given target bit-rate and a given input light field, the PSNR performance of the compression scheme depends on two key parameters: the rank k of the approximation and the HEVC quantization parameter (QP). To automatically select the best pair of parameters (k, QP), we train a model represented by a function f of a set of input features:

$$(k, \mathrm{QP}) = f(\text{LF features}, \text{Target bitrate}) \tag{15}$$

To generate training data labels, we encode at first several light fields with different values of k and QP, and then labels are extracted by taking the (k, QP) pairs only corresponding to the data points on the envelope of the PSNR-rate points.

1) Feature Space: The parameter prediction is regarded as a classification problem with the following input features:

1) Disparity indicators of original light field:
   a) proportion of singular values of the matrix I which contain at least 95% of the energy of I;
   b) decay rate of singular values of the matrix I which is defined as the ratio between the first and the second singular value.
2) Disparity indicators of aligned light field: same indicators as above for the matrix $I \circ h$.

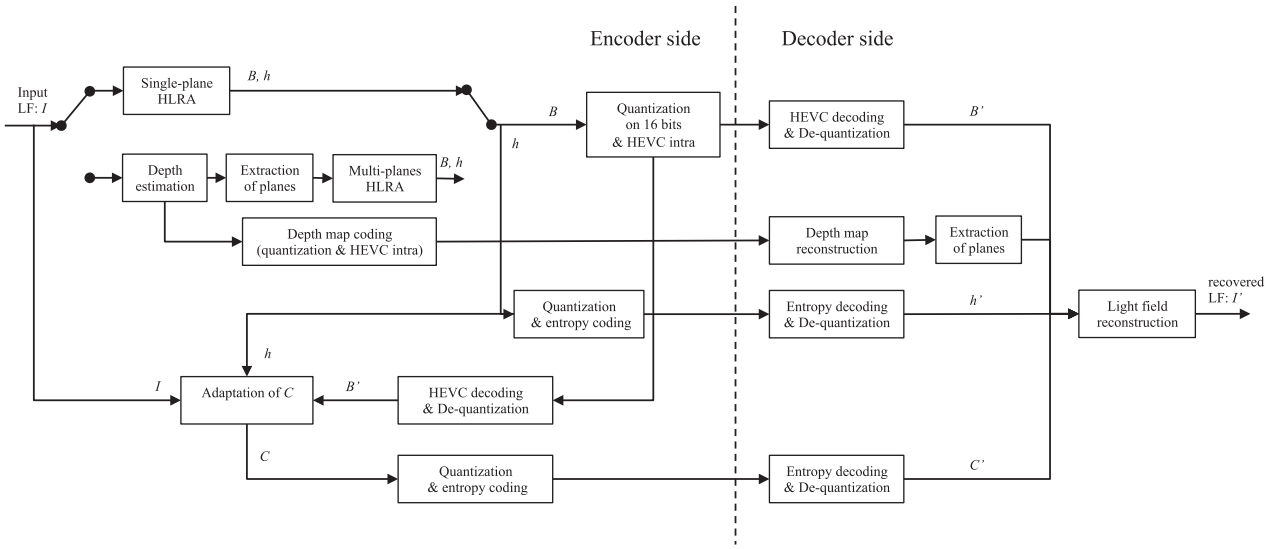


Fig. 5. Coding and decoding scheme overview.

3) Texture indicators: same indicators as above computed on the matrix in which each column is a vectorized version of each 8×8 block of the central view.
4) Bitrate of the encoded light field for a certain pair (k, QP). This feature gives some indication of the bitrate range for the input light field. In the experiments, we take $k = 15$, QP $= 14$ and encode the original LF (without alignment for a fast computation). Note that it requires nevertheless a supplementary encoding pass.
5) Target bitrate.

2) Decision Trees: The model input features are a mixture of continuous (e.g., bitrates and decay rate of singular values) and discrete (e.g., proportion of singular values keeping 95% of the energy) variables. Contrary to SVM or Logistic Regression which are only efficient with continuous inputs, Decision Trees (DT) are good candidates for dealing with data of different nature and do not need variable scaling. DT is also known to be robust to noisy data (outliers). Among models of DT ensemble, we use Random Forest [23] as classifier. The number of trees is 150, which is tuned by 10-fold cross-validation with our dataset.

3) Chain-Based Multi-Output Classification: The (k, QP) prediction task can be considered as a problem of Multi-output classification (MOC), a supervised learning problem where an instance is associated with a set of discrete labels, rather than with a single label. A classical way is to predict separately each label with a different classifier by assuming that these labels are independent. In our case, however, k and QP are strongly correlated. In order to improve the MOC performance, we model the label dependencies by a modified Classifier Chain (CC) at the expense of an increased computational cost.

As in [52], a CC model involves several classifiers, which are linked along a chain where each classifier deals with a classification problem associated with a different label. The predictions of the different classifiers are cascaded as additional features. In other words, the feature space of each link is extended with the label associations of all previous links. We have observed that a competitive chain scheme is experimentally better than simple unidirectional chains. In such a scheme, the values of k and QP are at first separately predicted by two independent Random Forests, each taking the feature space defined in Section VI-B1. We then choose the prediction (k or QP) for which the classifier gets a higher probability and add it into the new feature space. A third Random Forest is then employed to predict the other label with the augmented feature space.

VII. Performance Analysis

For simulations, we consider real light fields captured by plenoptic cameras using an array of micro-lenses, coming from different sources: 1/- the INRIA dataset [24] which contains LFs captured either by a first generation Lytro camera (63 LFs, 11×11 views of 379×379 pixels per LF) or a second generation Lytro Illum camera (46 LFs, 15×15 views of 625×434 pixels per LF); 2/- the ICME 2016 Grand Challenge dataset [25] containing 12 Lytro Illum LFs.

Lytro LFs are decoded by the Matlab Light Field Toolbox v0.4 [53]. With the INRIA dataset, we only consider the 9×9 central views in order to alleviate the strong vignetting and distortion problems on the peripheral views, which comparatively impact more the performance of the HEVC-based reference schemes, e.g., [12], [13]. Note, however, that there are still variations of light intensity, to a lesser extent, in the truncated light fields. With the ICME 2016 Grand Challenge dataset, we take 13×13 central views as defined by the challenge testing conditions. The test light fields in this section are shown in Figs. 6 and 7.

For the INRIA dataset, the method in [49] has been used to estimate a depth map of the central view in order to test our method with multiple homographies per view as described in Section V. For the ICME 2016 Grand Challenge dataset, the provided depth maps have been used.

In experiments, the bitrate and PSNR are given for the luminance component. The PSNR is derived from the MSE (mean square error) computed on the whole light field.



Fig. 6. Test light fields in INRIA dataset. First line: Real Light Fields captured by a Lytro 1G camera (From left to right: TotoroWaterfall, Beers, Flower and TapeMeasure); Second line: Real Light Fields captured by a Lytro Illum camera (From left to right: Fruits, Bench, BouquetFlower1 and Toys).

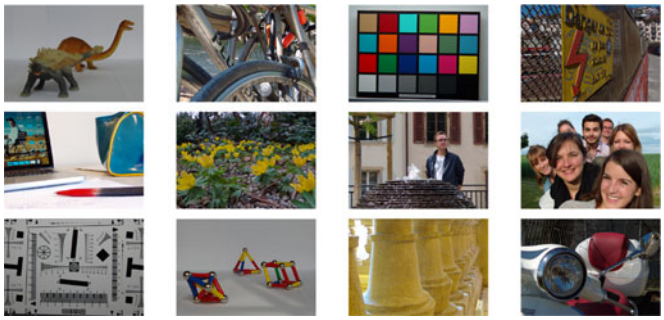


Fig. 7. Thumbnails of the ICME 2016 Grand Challenge dataset. From left to right, First line: Ankylosaurus_&_Diplodocus_1, Bikes, Color_Chart_1, Danger_de_Mort; Second line: Desktop, Flowers, Fountain, Friends; Third line: ISO_Chart_12, Magnets, Stone_Pillars_Outside, Vespa.

TABLE I
PSNR OBTAINED WITH DIFFERENT VALUES OF RANK k FOR HOMOGRAPHY OPTIMIZATION AND LOW RANK APPROXIMATION

aligning rank	approximation rank	PSNR (dB)	
		TotoroWaterfall	Toys
5	5	34.44	34.67
30	30	45.98	41.94
30	5	33.04	33.12
5	30	40.23	38.33

A. Analysis of HLRA-Based Compression

1) Joint Homography and Low Rank Optimization: We first assess the benefit of the joint optimization of the homographies and of the B and C matrices. Table I shows the PSNR obtained with different values of rank k for homography search and low rank approximation. The same value of k in both columns means that the same rank is used for computing the homographies and the transmitted matrices B and C. By comparing the first and the third row on one hand, and the second and fourth row on the other hand, for both light fields, one can see that for a given approximation rank, a joint optimization of homographies and of the approximation brings a significant gain.

2) Alignment Gain: Fig. 8 shows the interest of view alignment for compression for three LFs: "TotoroWaterfall", "Toys" and "Ankylosaurus_&_Diplodocus_1". The gain of applying homographies to light field views is mostly significant at low bitrates and with a low approximation rank. Note that for plenoptic cameras using arrays of micro-lenses, disparity across views is relatively limited, hence alignment by one homography per view is usually sufficient to satisfy low rank assumption. When the disparity significantly varies across the scene (e.g., "TotoroWaterfall"), multiple homographies can further improve the compression performance, despite the additional bitrate cost for transmitting the depth map.

3) Accuracy of the Parameters Prediction Model: Fig. 8 also suggests that the PSNR-rate performance depends on the values of the approximation rank k and of the HEVC coder quantization parameter (QP). For example, it appears that a smaller approximation rank k is preferred at low bit-rate. It can also be observed that the best pair of parameters (k, QP) depends on the input light field, hence the need to model the relationship between the target bit-rate, input light field features and the best values for these parameters.

The model proposed in Section VI-B for predicting the best pair of (k, QP) has been trained using light fields in the INRIA dataset [24]. The dataset contains both indoor and outdoor captures and the light fields are taken with variable focal length. The test content is not similar to the training content. For each type of camera (Lytro first generation or Illum), we chose 4 light fields for the test (*cf.* Fig. 6). The remaining light fields are used for training and they correspond to different types of scenes. Training and test LFs are available for download on the website.[1] For each LF, compression is performed with the HLRA scheme for a combination of different values of k and QP ($k \in \{5, 15, 30\}$ and QP $\in \{2, 6, 10, 14, 20, 26, 38\}$). Finally, training data labels consist of (k, QP) pairs only corresponding to the data points on the envelope of the PSNR-rate points.

Table II shows the exact matching rates of the competitive classifier chain (competitive CC) compared to other classification methods: 1/- independent classifiers: k and QP are independently predicted; 2/- CC k –> QP: classifier chain with k predicted before QP; 3/- CC QP –> k: classifier chain with QP predicted before k. Note that in competitive CC scheme, with two labels to classify, there are four classifiers to train and three of them are employed at the test time, whereas in a classical CC model, the number of classifiers is two at both the training and test phase. Fig. 9 shows in solid lines the true envelope after testing all possible (k, QP) pairs, and in dashed lines the PSNR-rate curves corresponding to the predicted (k, QP) pairs. Both one (black curves) and two homographies per view (red curves) are investigated. Although the exact match of (k, QP) values is not achieved for about $1/3$ of cases (Table II), we observe that the predicted PSNR-rate curve is very close to the true envelope.

In order to avoid multiple trainings for different numbers of homographies, the training has been done with a single homography. However, one can see in Fig. 9 that the prediction model learned on light fields for one homography remains valid for multiple homographies.

4) Bitrate Cost Percentage Analysis: The percentage of the bitrate allocated to each element is detailed in Table III. A dominant part of bits are allocated to encoding the matrix B. The homography parameters in h and the depth map D require a fixed cost that does not vary in function of the target bitrate. As

[1] https://www.irisa.fr/temics/demos/lightField/LowRank2/datasets/datasets.html

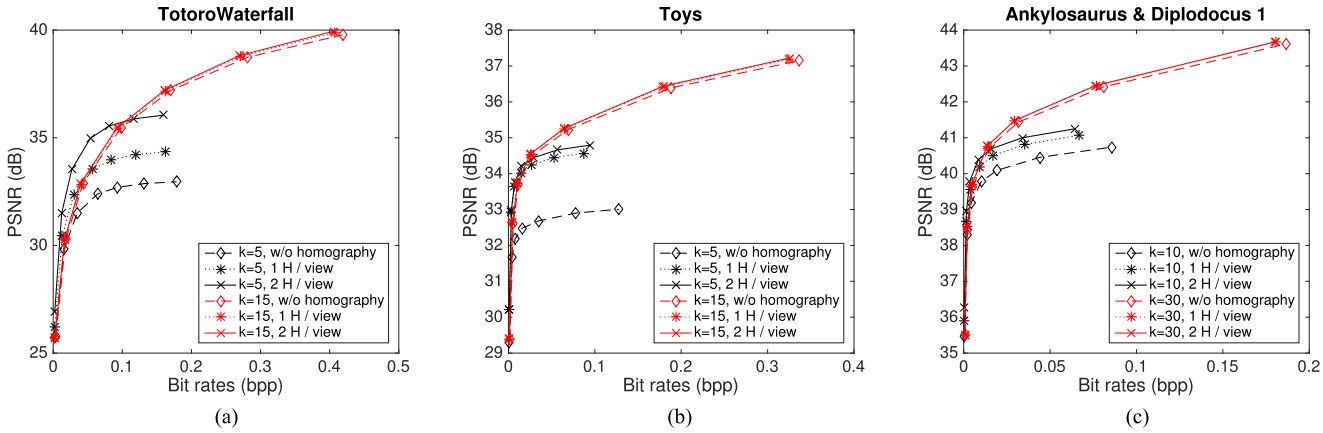


Fig. 8. PSNR-rate performance of HLRA: with (1 or 2 homographies per view) or without alignment of light field views.

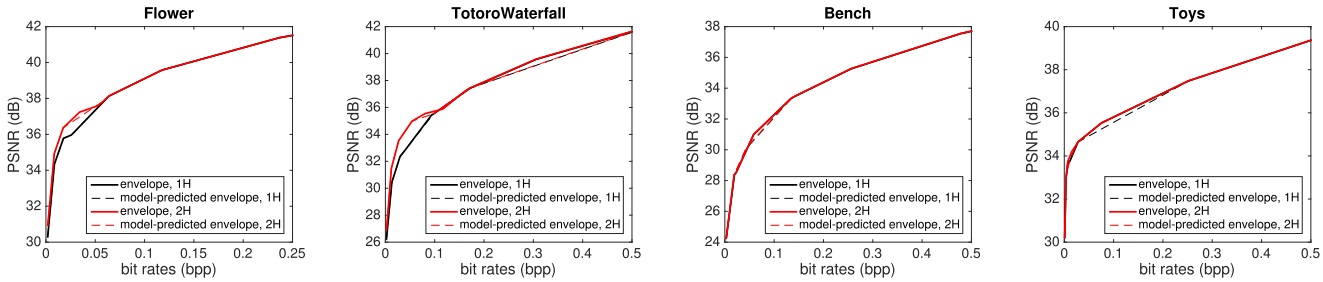


Fig. 9. Performance evaluation of model-based parameter prediction for 4 of the test LFs. The PSNR-rate curves corresponding to model-predicted coding parameters are shown by dashed lines, and the envelope obtained by all possible combinations of coding parameters are shown by solid lines. Both one (black curves) and two homographies per view (red curves) are investigated.

TABLE II
EXACT MATCHING RATE OF DIFFERENT CLASSIFICATION SCHEMES

	independent classifiers	CC k –> QP	CC QP –> k	competitive CC
Lytro 1G	64.2%	65.7%	66.1%	67.1%
Lytro Illum	59.2%	62.0%	62.0%	64.1%

a consequence, the percentage of their cost becomes negligible at high bitrates.

B. Comparative Assessment of Compression Performance

We assess the compression performance obtained with the homography-based low rank approximation against two schemes: direct encoding of the views as a pseudo-video sequence according to a lozenge order (HEVC-lozenge) [12] and according to the scanning order proposed in [13] (HEVC-pseudo). In simulations, the base QPs of HEVC-pseudo are set to $\text{QP}_B = 8$, 14, 20, 26, 32 and 38, and the views at hierarchical layers 2, 3, 4, 5, 6 respectively have QPs equal to $\text{QP}_B + 8, \text{QP}_B + 9, \text{QP}_B + 10, \text{QP}_B + 11$ and $\text{QP}_B + 12$, as described in [13]. For HEVC-lozenge, the base QPs are set to 20, 26, 32, 38 and a GOP of 4 is used. The HEVC version used in the tests is HM-16.10.

In Figs. 10 and 11, both HLRA with 1H (one homography per view) and HLRA with 2H (two homographies per view) are investigated against HEVC-lozenge and HEVC-pseudo. Table IV, gives the PSNR gain (using the Bjontegaard measure) of the proposed method with one or two homographies in comparison with the method HEVC-pseudo [13]. The reference considered for computing the Bjontegaard measure is HEVC-lozenge [12]. Furthermore, the approximation error of the center view obtained with HEVC-lozenge, HEVC-pseudo and HLRA is given in Fig. 12 for comparison.

Substantial gains in favor of HLRA schemes are observed for most of the test light fields. In Table IV, in most cases, best results are obtained with HLRA using two homographies per view, though on average, one homography per view is sufficient to outperform the HEVC-pseudo scheme. HLRA with two homographies are especially interesting when the scene contains several depth planes, such as “Desktop”, “TotoroWaterfall” and “Flower”. Note that there is strictly only one depth plane in “Color_Chart_1” and “ISO_Chart_12”, and the inaccuracy in depth map estimation explains the degradation when using 2 homographies per view for these two LFs.

For real world LFs captured by plenoptic cameras, global variations of light intensity are present between views. Although this degrades the performance of HEVC inter coding, our compression scheme is little affected since the rank of a matrix remains constant when its columns are multiplied by different factors.

TABLE III
PERCENTAGE OF THE BITRATE ALLOCATED TO EACH ELEMENT

Datasets	Nb. of H per view	Traget bitrates (bpp)	B	C	h	D
ICME Grand Challenge	1H	4.5×10^{-3} (k=10, QP=26)	81.9%	11.5%	6.6%	-
	1H	9.8×10^{-2} (k=60, QP=6)	95.7%	4.0%	0.3%	-
	2H	5.1×10^{-3} (k=10, QP=26)	67.6%	9.5%	11.1%	11.8%
	2H	9.9×10^{-2} (k=60, QP=6)	94.8%	4.0%	0.5%	0.7%
INRIA Lytro Dataset	1H	1.0×10^{-2} (k=5, QP=26)	92.4%	3.3%	4.3%	-
	1H	3.0×10^{-1} (k=60, QP=6)	99.1%	0.8%	0.1%	-
	2H	1.2×10^{-2} (k=5, QP=26)	77.5%	2.7%	7.5 %	12.3%
	2H	3.0×10^{-1} (k=30, QP=6)	98.4%	0.8%	0.3%	0.5%

The results are averaged over test light fields in the ICME 2016 Grand Challenge dataset and INRIA Lytro Dataset, respectively.

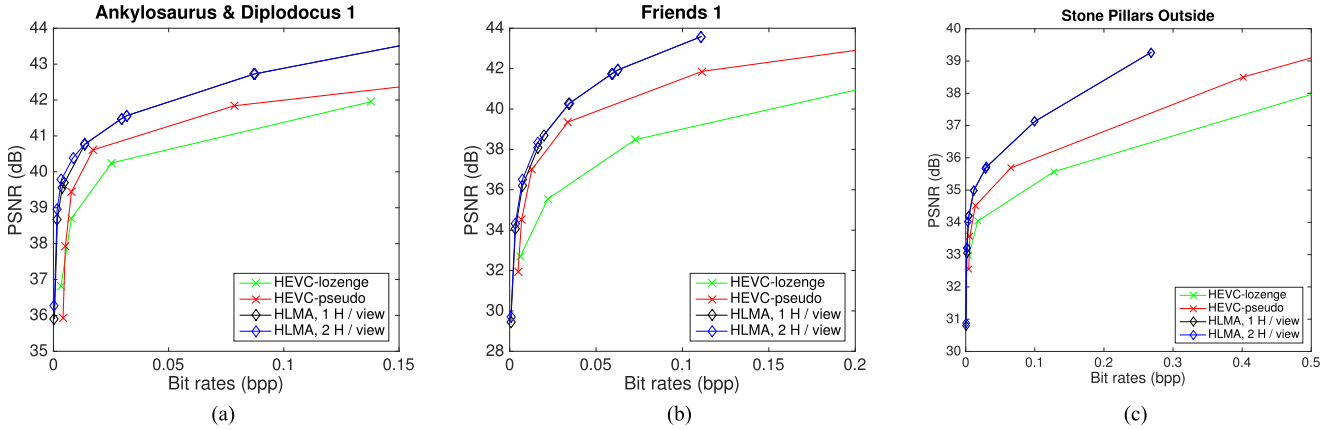


Fig. 10. PSNR-rate performance comparisons with three images from the ICME 2016 Grand Challenge dataset.

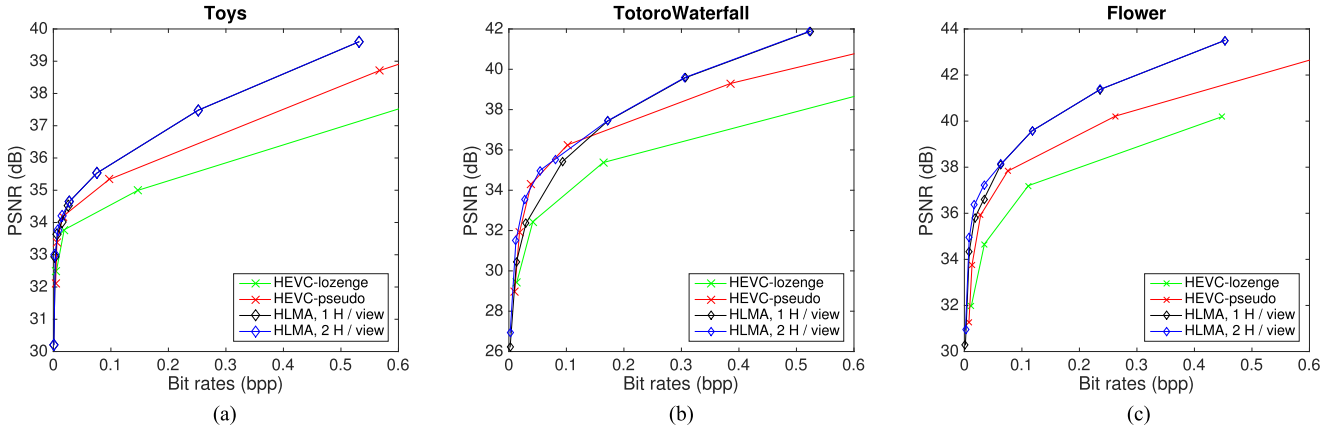


Fig. 11. PSNR-rate performance comparisons with three images from the INRIA dataset.

Note that the PSNR of the central views (the 7×7 views in the center of the light field) is quite stable. However, the PSNR of the reconstructed views at the periphery of the light field varies. The gap can be up to 2–3 dB for real LFs between the central views and the views at the periphery. This variation is due to 2 reasons: 1/- a peripheral view requires a more significant homography transformation in order to align with other views, and the error due to forward and inverse warping is consequently more important. 2/- peripheral views suffer more severely from noise and distortion than central views. These artifacts are removed by low rank approximation, which causes the drop of PSNR, but not necessarily the degradation of visual quality. One may refer to the website[2] to see the visual quality of reconstructed views. Furthermore, this variation of PSNR

[2]https://www.irisa.fr/temics/demos/lightField/LowRank2/LRcompression.html

TABLE IV
BD-PSNR GAINS WITH RESPECT TO HEVC-LOZENGE SCHEME [12]

Datasets	LF Names	HEVC-pseudo [12]	HLRA with 1H	HLRA with 2H
ICME Grand Challenge $13 \times 13 \times 625 \times 434$	Bikes	2.53	2.87	**2.91**
	Danger_de_Mort	2.74	**3.60**	3.46
	Flowers	3.16	3.22	**3.33**
	Stone_Pillars_Outside	0.70	1.47	**1.48**
	Vespa	1.79	1.79	**2.00**
	Ankylosaurus_&_Diplodocus_1	0.58	1.30	**1.35**
	Desktop	2.00	1.95	**2.14**
	Magnets_1	0.56	1.48	**1.49**
	Fountain_&_Vincent_2	**2.08**	1.51	1.82
	Friends_1	2.40	3.37	**3.47**
	Color_Chart_1	0.84	**1.73**	1.67
	ISO_Chart_12	**1.97**	1.90	1.77
	Average	1.78	2.18	**2.24**
INRIA Lytro Illum $9 \times 9 \times 625 \times 434$	BouquetFlower1	1.78	2.12	**2.13**
	Toys	0.70	1.55	**1.56**
	Bench	2.47	2.76	**2.77**
	Fruits	**2.72**	1.92	2.13
	Average	1.92	2.09	**2.15**
INRIA Lytro 1G $9 \times 9 \times 379 \times 379$	Beers	1.55	2.24	**2.26**
	TotoroWaterfall	1.85	1.50	**2.08**
	Flower	1.47	2.36	**2.52**
	TapeMeasure	1.26	1.63	**1.69**
	Average	1.53	1.93	**2.14**
Overall average		1.76	2.11	**2.20**

The gains are shown for the HEVC-pseudo [13] and for our HLRA scheme with one or two homographies per view.

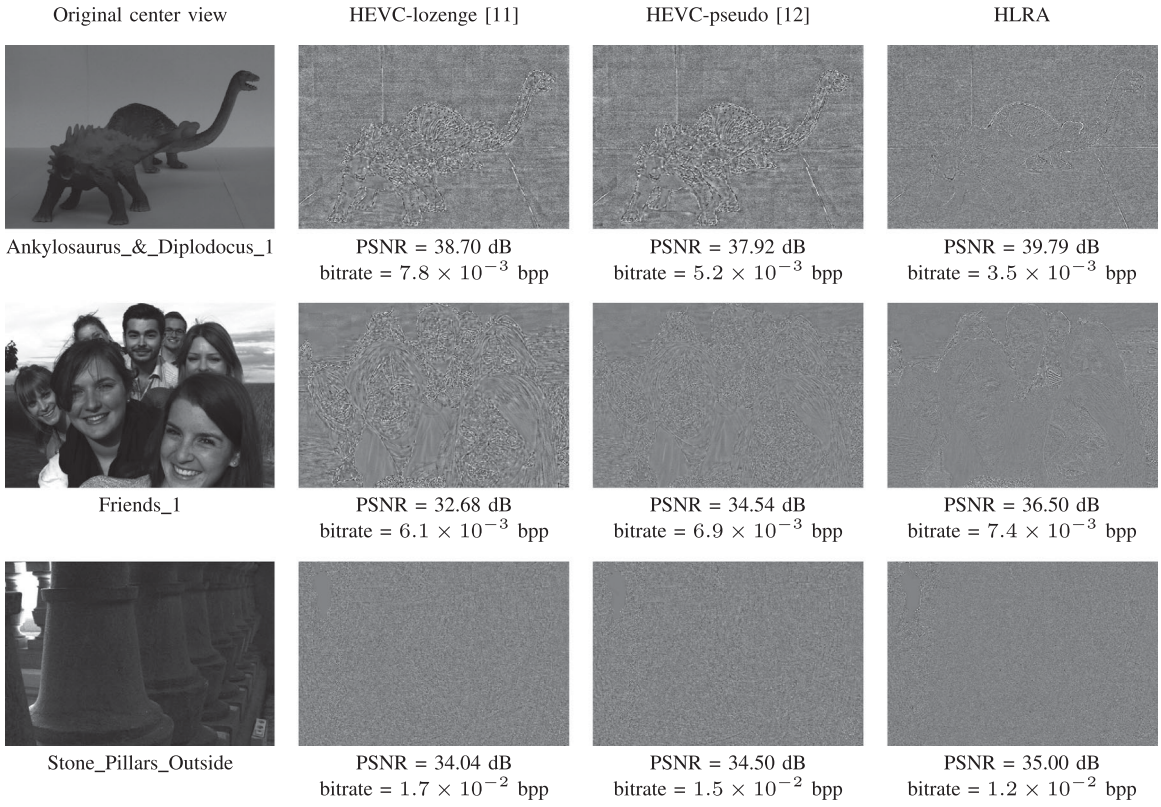


Fig. 12. The approximation error of the center view obtained with HEVC-lozenge, HEVC-pseudo and HLRA compression schemes. Similar bitrates are chosen for the three schemes.

TABLE V
RUNTIMES

Datasets		HLRA			HEVC-pseudo [12]
		Searching H		Encoding B	
		Nb. iters	t (min)	t (min)	t (min)
ICME Grand Challenge $13 \times 13 \times 625 \times 434$	$k = 10$	29.9	12.5	0.6	17.0
	$k = 30$	7.3	3.2	1.7	
	$k = 60$	2.7	2.7	3.3	
INRIA Lytro Illum $9 \times 9 \times 625 \times 434$	$k = 5$	40.0	8.9	0.4	8.3
	$k = 15$	7.8	1.8	0.9	
	$k = 30$	6.3	1.5	1.7	
INRIA Lytro 1G $9 \times 9 \times 379 \times 379$	$k = 5$	20.8	2.2	0.2	5.0
	$k = 15$	5.0	0.6	0.5	
	$k = 30$	4.8	0.5	0.9	

The results are averaged over the test light fields in different datasets. The number of required iterations (Nb. iters) and runtime (t) are detailed. The runtime is measured at $\text{QP}_B = 20$ both for HLRA and HEVC-pseudo [13].

between views is compensated in the two layer scheme presented in Section VIII by transmitting the residue.

C. Computational Complexity Analysis

The computational complexity of HLRA scheme mainly resides in two parts: finding the set of homographies to align the light field views, and encoding the matrix B. Searching for homographies is an iterative procedure, each iteration containing successive steps: 1/- warping involves multiplying each pixel coordinate in each light field view $[x, y, 1]$ by a 3×3 matrix, the complexity being $\mathcal{O}(nm)$; 2/- computing the Jacobian J_i for $i \in [|1, n|]$ implies $\mathcal{O}(nm)$ operations; 3/- the complexity for the SVD of the matrix $I' = I \circ h$ is $\mathcal{O}(n^2 m)$; 4/- and finally for $i \in [|1, n|]$, computing Δh_i involves computing the pseudo-inverse of the Jacobian matrix J_i of size $m \times 8$, which has a complexity of $\mathcal{O}(m)$, followed by a matrix multiplication with complexity $\mathcal{O}(m)$. Overall, each iteration requires $\mathcal{O}(n^2 m)$ arithmetic operations.

In our experiments, we consider that the algorithm has converged when the PSNR gain is less than 0.002 dB between two successive iterations. Logically, an approximation with smaller rank requires more iterations to converge. In fact, in the extreme case where the approximation rank is equal to the number of light field views, no alignment (hence zero iteration) is needed and the decomposition in B and C is reduced to a simple SVD. The encoding time of the matrix B is proportional to the number of columns (the approximation rank), since each column is intra coded by HEVC. Considering their small size, the encoding time of C, h and the depth map is negligible.

The number of iterations and the consumed time for finding homographies and encoding B are detailed in Table V. Simulations have been carried out with a Macbook Pro with a 2.8 GHz Intel Core i7 processor. In spite of the additional cost for the iterative alignment, HLRA consumes less time in total than the HEVC-pseudo scheme because the intra coding of the columns of B is much faster than inter coding of all the views in HEVC-pseudo. Note that our implementation of the iterative alignment is written in Matlab and could be further optimized in the future.

On the decoder side, only k images need to be decoded with HEVC intra, and the remaining steps are the matrix multiplication BC (complexity $\mathcal{O}(mnk)$) and inverse warpings (complexity $\mathcal{O}(mn)$). Note that the decoding process is not iterative since the homographies and the matrices B and C are directly transmitted to the decoder.

TABLE VI
BD-PSNR GAINS (WITH RESPECT TO HEVC-LOZENGE SCHEME [12]) EVALUATION FOR SYNTHETIC LIGHT FIELDS IN HCI DATASET

LF Names	HEVC-pseudo [13]	HLRA
Buddha	4.72	1.00
Butterfly	2.52	0.17
MonasRoom	3.90	1.08
StillLife	4.37	−4.28

The gains are shown for the HEVC-pseudo [13] and for our HLRA schemes.

D. Limitations of the Method

For synthetic light fields in HCI dataset [54], the HLRA scheme performs in general better than HEVC-lozenge but worse than HEVC-pseudo in terms of PSNR-rate performance (*cf.* Table VI). Two main reasons may explain this degradation. First, while the baseline of the light fields captured with plenoptic cameras is limited by the aperture size of the camera, there is no such limitation with synthetic light fields which may then have much higher disparities between views. In these conditions, global homography projections less accurately compensate for the inter-view disparities than the block-matching of HEVC inter. With large baselines, the first columns of the matrix B contain considerably more information, even after view alignment, and are therefore more expensive to encode. This is the reason why, in Table VI, the gap between our method and HEVC-pseudo is more important for light fields with larger baselines ("StillLife" and "Buddha") than for those with smaller baselines ("Butterfly" and "MonasRoom"). Secondly, synthetic light fields are free of imperfections such as noise and variations of light intensity between views. The real world LF imperfection degrades considerably the performance of HEVC inter coding while HLRA compression scheme is little affected. In fact, a global change of light intensity in one view simply leads to

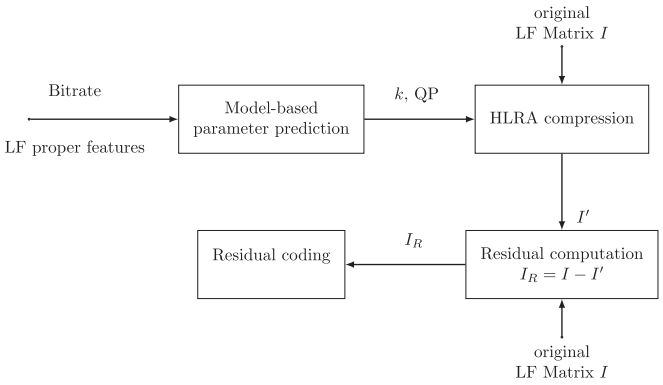


Fig. 13. Scalable light field coding chain.

multiplying the corresponding coefficients in the matrix C by different factors, while the noise is mostly dropped by the low rank model.

VIII. Scalable Light Field Coding

Thanks to the matrix factorization used in our approach, our coding scheme naturally presents an interesting scalability property. Although the encoder transmits a fixed number of columns of the matrix B corresponding to the rank k, the decoder can choose to decode less than k columns. Since the matrices B and C are obtained by a singular value decomposition which sorts the singular values and their corresponding singular vectors, the first columns of B contain most of the energy of the light field signal. Therefore, a fast approximation of the encoded light field can already be performed on the decoder side by decoding only those first columns. The decoder may then progressively refine the low rank approximation by decoding additional columns.

However, in addition to the errors caused by the low rank approximation, we have identified two other types of errors either caused by the quantization in the HEVC compression, or by the forward and inverse homography warping. These errors are not corrected by encoding additional columns of B. Therefore, we propose a scalable extension of our method where the residual between the original and the decoded light field is encoded as an enhancement layer, as illustrated in Fig. 13.

For our experiment, the base layer is computed by encoding the original light field with HLRA using a single homography. For the residual layer, four coding schemes are tested:

1) HEVC-lozenge encoding [12].
2) HEVC-pseudo encoding [13].
3) HLRA without alignment (only BC factorization).
4) HLRA with one homography.

Fig. 14 shows the PSNR-rate performance of the scalable light field coding. For comparing the different residual layer coding schemes, the base layer is encoded at a fixed target bitrate of 0.01 bpp. The corresponding k and QP parameters are automatically predicted by our classification model decribed in Section VI-B. For the HLRA encoding of the residual (either with or without homography alignment), results are shown for different values of the approximation rank k_r. Each curve is

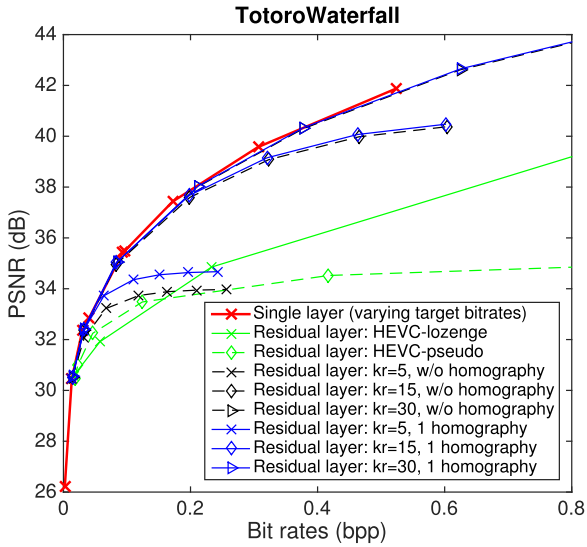


Fig. 14. Scalable coding performance for a base layer encoded with the single homography HLRA. The red curve shows the single layer result at varying target bitrates. The other curves show the scalable performance for different residual encoding methods, where the base layer is encoded at a target bitrate of 0.01 bpp ($k = 5$, QP$= 26$).

generated by varying the QP parameter in the HEVC encoding of the residual B matrix.

We first note that all the variants of our HLRA scheme perform significantly better than the HEVC-lozenge and HEVC-pseudo schemes applied to the residual layer.

Unlike the base layer (*cf.* Fig. 8), aligning the residual and encoding with a low rank k_r does not significantly improve the results at low or medium bitrates compared to a higher rank encoding. The best performance at any bitrate is then obtained by choosing the highest rank k_r. In this case, homography alignment only brings very little gains which may not justify the added complexity of the optimization procedure for determining homographies for the residue. The BC factorization computed by a single SVD step is then sufficient for the residual matrix encoding.

Finally, in comparison to the single layer approach (red curve), encoding a residual layer with a high approximation rank ($k_r = 30$) only results in a negligible loss compared to single layer HEVC-based coding while providing scalability. Note that similarly to the base layer, the residual layer encoded with the BC factorization can be decoded by progressive refinement when decoding the successive columns of the B matrix.

IX. Conclusion

In this paper, we have proposed a new compression scheme for light field images. In our method, each view of the light field is first warped using either one global homography, or a model consisting of a homography per depth plane with a smooth transition between depth planes. Considering the matrix formed by concatenating each warped and vectorized view, a joint optimization of homography parameters and low rank matrix approximation is performed. For an approximation rank k, the resulting rank k matrix can be factorized as the product

of a matrix containing k basis vectors and a smaller matrix of coefficients. Our method then consists in encoding each of the k basis vectors using HEVC intra, along with the coefficients and homography parameters. In the case of multiple depth planes, the depth map must be encoded as well using HEVC intra. For most tested light fields, experimental results show substantial performance gains compared to the state-of-the-art methods which encode the views as a pseudo-sequence with HEVC inter. The method is particularly well suited for real light fields captured with plenoptic cameras which have limited disparity, but may contain imperfections such as variations of light intensity across views.

Our method is dependent on two encoding parameters: the rank k and the QP parameter in HEVC. Therefore, we have also proposed a prediction scheme for determining the best couple (k, QP) as a function of a target bitrate and additional features computed on the input light field. Our experiments show that the predicted parameters always result in close to optimal coding performance. Furthermore, the model remains valid in broader conditions than what it was trained for (i.e., multiple homographies).

Finally, a scalable extension has been proposed where a residual layer is encoded. We have shown that homography warping is not necessary for the residual. A simpler coding scheme only based on matrix factorization is sufficient and it substantially outperforms the HEVC inter encoding of the pseudo-sequence of view residuals.

ACKNOWLEDGMENT

The authors would like to thank Dr. Dong Liu for his kindness to share the source code of the ICME 2016 Grand Challenge paper "Pseudo-sequence-based light field image compression" [13].

REFERENCES

[1] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, Jul. 2005.

[2] R. Ng, "Light field photography," Ph.D. dissertation, Stanford Univ., Stanford, CA, USA, 2006.

[3] T. Georgiev, G. Chunev, and A. Lumsdaine, "Super-resolution with the focused plenoptic camera," *Proc. SPIE*, vol. 7873, 2011, Art. no. 78730X.

[4] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Annu. Conf. Comput. Graph. Interactive Techn.*, 1996, pp. 31–42.

[5] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.

[6] P. Lalonde and A. Fournier, "Interactive rendering of wavelet projected light fields," in *Proc. Int. Conf. Graph. Interface*, 1999, pp. 107–114.

[7] I. Peter and W. Straßer, "The wavelet stream—progressive transmission of compressed light field data," in *Proc. IEEE Vis. 1999 Late Breaking Hot Topics*, 1999, pp. 69–72.

[8] M. Magnor and B. Girod, "Data compression for light-field rendering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 3, pp. 338–343, Apr. 2000.

[9] C.-L. Chang, X. Zhu, P. Ramanathan, and B. Girod, "Light field compression using disparity-compensated lifting and shape adaptation," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 793–806, Apr. 2006.

[10] M. Magnor, A. Endmann, and B. Girod, "Progressive compression and rendering of light fields," in *Proc. Vis., Model. Vis.*, 2000, pp. 199–203.

[11] D. Lelescu and F. Bossen, "Representation and coding of light field data," *Graph. Models*, vol. 66, no. 4, pp. 203–225, Jul. 2004.

[12] M. Rizkallah, T. Maugey, C. Yaacoub, and C. Guillemot, "Impact of light field compression on focus stack and extended focus images," in *Proc. 24th Eur. Signal Process. Conf.*, Aug. 2016, pp. 898–902.

[13] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2016.

[14] X. Jiang, M. L. Pendu, R. A. Farrugia, S. S. Hemami, and C. Guillemot, "Homography-based low rank approximation of light fields for compression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 1313–1317.

[15] C. Conti, P. Nunes, and L. D. Soares, "New HEVC prediction modes for 3D holoscopic video coding," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1325–1328.

[16] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Efficient intra prediction scheme for light field image compression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Florence, Italy, May 2014, pp. 539–543.

[17] C. Perra and P. Assuncao, "High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2016.

[18] C. Conti, P. Nunes, and L. D. Soares, "HEVC-based light field image coding with bi-predicted self-similarity compensation," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2016.

[19] R. Monteiro, L. Lucas, C. Conti, P. Nunes, N. Rodrigues, S. Faria, C. Pagliari, E. da Silva, and L. Soares, "Light field HEVC-based image coding using locally linear embedding and self-similarity compensated prediction," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2016.

[20] Y. Li, R. Olsson, and M. Sjöström, "Compression of unfocused plenoptic images using a displacement intra prediction," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2016.

[21] T. Sakamoto, K. Kodama, and T. Hamamoto, "A study on efficient compression of multi-focus images for dense light-field reconstruction," in *Proc. Vis. Commun. Image Process.*, Nov. 2012, pp. 1–6.

[22] A. Levin and F. Durand, "Linear view synthesis using dimensionality gap light field prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1831–1838.

[23] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[24] "INRIA Lytro image dataset." [Online]. Available: https://www.irisa.fr/temics/demos/lightField/LowRank2/datasets/datasets.html

[25] "ICME 2016 Grand Challenge dataset." [Online]. Available: http://mmspg.epfl.ch/EPFL-light-field-image-dataset

[26] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. 23rd Annu. Conf. Comput. Graphics Interactive Techn.*, 1996, pp. 43–54.

[27] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, Mar. 2010.

[28] R. Meka, P. Jain, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 937–945.

[29] K. Lee and Y. Bresler, "ADMiRA: Atomic decomposition for minimum rank approximation," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4402–4416, Sep. 2010.

[30] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

[31] T. Zhou and D. Tao, "GoDec: Randomized low-rank & sparse matrix decomposition in noisy case," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 33–40.

[32] J. Tanner and K. Wei, "Normalized iterative hard thresholding for matrix completion," *SIAM J. Sci. Comput.*, vol. 35, no. 5, pp. S104–S125, 2013.

[33] Y. D. Kim and S. Choi, "Weighted nonnegative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2009, pp. 1541–1544.

[34] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Math. Program. Comput.*, vol. 4, no. 4, pp. 333–361, 2012.

[35] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," arXiv:1009.5055, Sep. 2010.

[36] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Commun. ACM*, vol. 55, no. 6, pp. 111–119, Jun. 2012.
[37] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM J. Optim.*, vol. 21, no. 2, pp. 572–596, 2011.
[38] J. Wright, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, p. 2080.
[39] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 763–770.
[40] C. Chen, J. Cai, W. Lin, and G. Shi, "Incremental low-rank and sparse decomposition for compressing videos captured by fixed cameras," *J. Vis. Commun. Image Represent.*, vol. 26, pp. 338–348, Jan. 2015.
[41] J. Hou, L. P. Chau, N. Magnenat-Thalmann, and Y. He, "Sparse low-rank matrix approximation for data compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 5, pp. 1043–1054, May 2017.
[42] C. Conti, L. D. Soares, and P. Nunes, "HEVC-based 3D holoscopic video coding using self-similarity compensated prediction," *Signal Process., Image Commun.*, vol. 42, pp. 59–78, Jan. 2016.
[43] L. Lucas, C. Conti, P. Nunes, L. Soares, N. Rodrigues, C. Pagliari, E. Silva, and S. Faria, "Locally linear embedding-based prediction for 3D holoscopic image coding using HEVC," in *Proc. 22nd Eur. Signal Process. Conf.*, Sep. 2014, pp. 11–15.
[44] T. Georgiev and A. Lumsdaine, "Focused plenoptic camera and rendering," *J. Electron. Imag.*, vol. 19, no. 2, Apr. 2010, Art. no. 021106.
[45] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Scalable coding of plenoptic images by using a sparse set and disparities," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 80–91, Jan. 2016.
[46] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1027–1034.
[47] S. Kundu, "Light field compression using homography and 2D warping," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 1349–1352.
[48] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.
[49] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 673–680.
[50] H. G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. W. Tai, and I. S. Kweon, "Accurate depth map estimation from a lenslet light field camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1547–1555.
[51] T. C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3487–3495.
[52] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, no. 3, pp. 333–359, 2011.
[53] D. Dansereau, "Light field toolbox for v0.4 by D. Dansereau," 2015. [Online]. Available: https://fr.mathworks.com/matlabcentral/fileexchange/49683-light-field-toolbox-v0-4
[54] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," in *Proc. Vis., Model. Vis. Workshop*, 2013, pp. 225–226.



Xiaoran Jiang received the Engineering degree in telecommunication and the Ph.D. degree in neural networks from the Institut Mines-Télécom (Télécom Bretagne), Brest, France, in 2010 and 2014, respectively. He is currently a Postdoctoral Researcher in the Institut National de Recherche en Informatique et en Automatique, Rennes, France. His research interests include signal and image processing, computer vision, and in particular low-rank approximation and view synthesis for light field images.



Mikaël Le Pendu received the Engineering degree from the Ecole Nationale Supérieure des Mines de Nantes, Nantes, France, in 2012, and the Ph.D. degree in computer science from the University of Rennes 1, Rennes, France, in 2016. His Ph.D. studies were conducted in conjunction between the Institut National de Recherche en Informatique et en Automatique (INRIA) and Technicolor in Rennes, France, and addressed the compression of high-dynamic range video content with a particular focus on backward compatibility. He is a Postdoctoral Researcher at INRIA. His current research focuses on light field imaging and covers the full processing chain from capture to compression, including editing tasks.



Reuben A. Farrugia received the first degree in electrical engineering and the Ph.D. degree, both from the University of Malta, Msida, Malta, in 2004 and 2009, respectively. In January 2008, he was appointed Assistant Lecturer with the same department and is currently a Senior Lecturer. He has been in technical and organizational committees of several national and international conferences. In particular, he was the General-Chair of the IEEE International Workshop on Biometrics and Forensics and the Technical Programme Co-Chair of the IEEE Visual Communications and Image Processing in 2014. He has been contributing as a Reviewer of several journals and conferences, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO AND TECHNOLOGY, and the IEEE TRANSACTIONS ON MULTIMEDIA. On September 2013, he was appointed as the National Contact Point of the European Association of Biometrics.



Christine Guillemot (F'13) received the Ph.D. degree from the Ecole Nationale Superieure des Telecommunications, Paris, France, and the Habilitation for Research Direction degree from the University of Rennes, Rennes, France. She is the Director of Research in the Institut National de Recherche en Informatique et en Automatique, Rennes, France, and the Head of a research team dealing with image and video modeling, processing, coding, and communication. From 1985 to October 1997, she was with the France Telecom, where she was involved in various projects in the area of image and video coding for TV, HDTV, and multimedia. From January 1990 to mid-1991, she was with the Bellcore, Piscataway, NJ, USA, as a Visiting Scientist. Her research interests include signal and image processing, and in particular 2-D and 3-D image and video processing for various problems (compression, superresolution, inpainting, and classification).

Dr. Guillemot was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING (from 2000 to 2003, and from 2014 to 2016), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (from 2004 to 2006), and the IEEE TRANSACTIONS ON SIGNAL PROCESSING (from 2007 to 2009). She was a Senior Member of the editorial board of the IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING (from 2013 to 2015) and is currently the Senior Area Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING and member of the steering committee of the IEEE TRANSACTIONS ON MULTIMEDIA.

Minimum Description Length Sparse Modeling and Region Merging for Lossless Plenoptic Image Compression

Petri Helin, Pekka Astola, *Student Member, IEEE*, Bhaskar Rao, *Fellow, IEEE*, and Ioan Tabus, *Senior Member, IEEE*

***Abstract*—This paper proposes a complete lossless compression method for exploiting the redundancy of rectified light-field data. The light-field data consist of an array of rectified subaperture images, called for short views, which are segmented into regions according to an optimized partition of the central view. Each region of a view is predictively encoded using a specifically designed sparse predictor, exploiting the smoothness of each color component in the current view, and the cross similarities with the other color components and already encoded neighbor views. The views are encoded sequentially, using a spiral scanning order, each view being predicted based on several similar neighbor views. The essential challenge for each predictor becomes choosing the most relevant regressors from a large number of possible regressors belonging to the neighbor views. The proposed solution here is to couple sparse predictor design and minimum description length (MDL) principle, where the data description length is measured by an implementable code length, optimized for a class of probability models. This paper introduces a region merging segmentation under the MDL criterion for partitioning the views into regions having their own specific sparse predictors. In experiments, several fast sparse design methods are considered. The proposed scheme is evaluated over a database of plenoptic images, achieving better lossless compression ratios than straightforward usage of standard image and video compression methods for the spiral sequence of views.**

***Index Terms*—Lossless compression, light-field coding, minimum description length segmentation, plenoptics, sparse prediction.**

I. Introduction

THE field of plenoptics emerged recently as a prominent research area, although its roots date back to 1908 [1]. The initial mostly academic optics studies have evolved theoretically and technologically to such extent that nowadays plenoptic cameras are made commercially available for industrial applications and even for the consumer market.

Manuscript received December 31, 2016; revised April 27, 2017 and July 4, 2017; accepted July 14, 2017. Date of publication August 9, 2017; date of current version October 23, 2017. The guest editor coordinating the review of this paper and approving it for publication was Prof. Yebin Liu. *(Corresponding author: Petri Helin.)*

P. Helin, P. Astola, and I. Tabus are with Tampere University of Technology, 33720, Tampere, Finland (e-mail: petri.helin@tut.fi; pekka.astola@tut.fi; ioan.tabus@tut.fi).

B. Rao is with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093 USA (e-mail: brao@ucsd.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2737967

Recently, the compressibility of captured light-fields has attracted attention in the scientific community as an important topic. Several studies of the lossy compression of images captured with the recent plenoptic camera technologies were published, see e.g. [2] and [3].

In view of the wide availability of the cameras, standardization efforts are taking place for the compression of plenoptic camera lenslet images, under the initiative JPEG-PLENO of JPEG standardization committee, and two challenges were taking place at major conferences, ICME 2016 and ICIP 2017. The main focus of the standardization call-for-proposal and of the challenges is lossy compression in a wide range of compression rates. Several publications were submitted in response to the challenges, [4]–[8] and a first comparative study of their results was published in [9]. The performance is illustrated in terms of improving the objective compression performance with respect to JPEG legacy compressor, or with respect to a straightforward usage of HEVC compressor, and in terms of the subjective quality of the resulting stack of subaperture images, as in [9]. Lossy compression effects on refocused photograhps was studied in [10].

The main topic of this paper is the lossless compression of a light-field data structure, which consists of a stack of subaperture images. The lossless compression, or near-lossless compression, is also one item in the goals of the standardization initiative, since in some applications, like medical imaging, only lossless compression is permitted due to legal reasons. In our study the main goal is to optimize the objective compression performance, and since the images are compressed without any losses there is no need of a subjective evaluation of compression effects.

Previous work in lossless compression of plenoptic images was done by Perra [11], who studied the compressibility of the raw lenslet image captured at the sensor of a plenoptic camera. We address here the compressibility of the *final* light-field data structure, resulting after the pipeline of processing the raw lenslet image using a typical light-field toolbox, which is an array of subaperture images, representing the views of the scene from different angular points. By encoding and transmitting this final processed information, the decoder itself does not need to perform the pipeline of transforming a raw lenslet image into a light-field data structure.

The light-field data structure is an $N \times N$-array of subaperture images, each representing a different angular view of the scene. The central view is obtained at the direction of the camera

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

axis, the other views being called side views, shifted both horizontally an vertically from the camera axis. We encode in this paper the views sequentially, starting from the central view and advancing in a spiral order through the side views, exploiting the views that were already encoded in the encoding of a new side view.

We presented a similar but simpler compression scheme in the conference paper [12]. This paper proposes a more refined method, building on the initial scheme and adding new features in the parts which affect most the performance, so that the new reported results are better than the ones in [12]. A detailed discussion of the differences to [12] is presented in Section VI-C, after presenting all details of the new method. In essence, the partitioning method is MDL based, being driven by the final compression performance, and secondly, the sparse predictor design is improved by considering alternative sparse design algorithms and using a more complete MDL cost formula for selection of the optimal sparsity mask.

In Section II, we first present the main elements of the encoding scheme, including partition propagation, linear model setup, encoding of coefficients and residuals, which are similar to [12], but now there is also cross-color prediction and the entropic encoding strategy is more refined. In Section III, we investigate multiple methods for sparse predictor design (instead of opting for only one as in [12]) and for selecting the final structure of the predictor by the MDL criterion. In Section IV, we introduce several ways of obtaining the regions that are used for grouping similar pixels under a single region for a dedicated predictor, and propose a parameterless region merging method for obtaining an MDL-optimal set of regions. Section V shows experimental results obtained with the presented method and Section VI discusses the algorithm complexity and connections to [12]. Finally, Section VII presents the conclusions.

II. Compression Architecture

A. Definitions and Notations

A light-field structure L_F is an array with five dimensions, where an element $L_F(i,j,x,y,c)$ represents the color c, at the pixels with coordinates (x,y), from the view image corresponding to the angular direction (i,j). The angular direction of a view is indexed by the horizontal and vertical shifts of the viewing direction, $(i,j) \in \{1,\ldots,N\}^2$. The size of the sub-aperture view is $\xi \times \eta$, hence $(x,y) \in \{1,\ldots,\xi\} \times \{1,\ldots,\eta\}$. The set of values of the color c is defined for simplicity of notation as $\{0,1,2\}$, but we denote by that the color space RGB, so $\{R,G,B\} \equiv \{0,1,2\}$. The experimental results in this paper are given for the case of Lytro Illum plenoptic camera, where $(\xi,\eta) = (434,625)$ and $N = 15$. An individual view in the light-field structure is an RGB color image A_{ij} with elements $A_{ij}(x,y,c)$. We call central view the view $A_{i_c j_c}$, all other views being side views. For Lytro Illum $i_c = j_c = 8$, hence A_{88} is the central view.

The partition $\mathcal{P}_{i_c j_c} = \{\Omega^0_{i_c j_c},\ldots,\Omega^{L-1}_{i_c j_c}\}$ of A_{i_c,j_c}, can be represented as a matrix $Z_{i_c j_c}$ of labels (with labels from 0 to $(L-1)$), where $\Omega^0_{i_c j_c}$ is the closest region and $\Omega^{L-1}_{i_c j_c}$ the farthest. The set of pixels of any view, is denoted $\Omega = \{1,\ldots,\xi\} \times \{1,\ldots,\eta\}$. The partition label image Z_{ij} for the view A_{ij} induces a partition $\mathcal{P}_{ij} = \{\Omega^0_{ij},\ldots,\Omega^{L-1}_{ij}\}$ of Ω by the relation $\Omega^k_{ij} = \{(x,y)|Z_{ij}(x,y) = k\}$. The encoding scheme encodes the side views in a spiral order, illustrated in Fig. 1, with the linear index for enconding order denoted $\zeta(i,j)$, e.g., $\zeta(8,8) = 1, \zeta(8,7) = 2, \ldots, \zeta(15,1) = 225$.

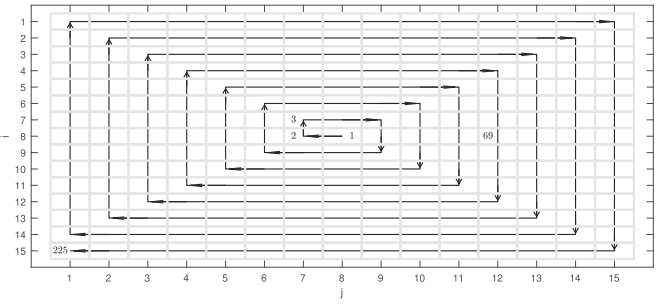


Fig. 1. The sequential processing order of the views is the shown spiral path starting from the center view $(8,8)$. The running index shown in a few selected views is denoted by ζ.

B. Partitions of Views Into Regions

A central role in the encoding scheme is played by the partition of each side view into regions. The central partition matrix $Z_{i_c j_c}$ in this work is designed by one of the methods described in Section IV. The label image Z_{ij} associated with the side view (i,j) is obtained with Algorithm 1 as explained below.

We assume fronto-planar regions, in order to simplify the process of finding the partitions induced in the side views by the partition in the central view. A region $\Omega^k_{i_c j_c}$ is displaced by δ^k_{ij} pixels in the horizontal direction and by μ^k_{ij} pixels in the vertical direction for defining the region Ω^k_{ij} in the partition of A_{ij}. For estimating the displacements we use the approximate relation at non-occluded pixels $A_{i_c j_c}(x,y) \approx A_{ij}(x+\delta^k_{ij}, y+\mu^k_{ij})$ and obtain the displacements as the solutions of the minimization

$$(\delta^k_{ij},\mu^k_{ij}) = \underset{\delta,\mu\in\Delta}{\arg\max} \sum_{x,y\in\Omega^k_{i_c j_c}} \sum_{c=0}^{2} \left(A_{i_c j_c}(x,y,c) - A_{ij}(x+\delta, y+\mu, c)\right)^2 . \quad (1)$$

The set of allowed displacements in this paper is $\Delta = \{-15,\ldots,16\}$, hence five bits are needed for representing each displacement.

The displacements for all 224 side views and regions are sent to the decoder, who receives also the central partition $Z_{i_c j_c}$ and is able to construct the side view partitions in the same way as the encoder, as described in Algorithm 1.

When propagating all regions $\Omega^0_{i_c j_c},\ldots,\Omega^{L-1}_{i_c j_c}$ according to the found set of displacements for one view A_{ij} there will be situations of conflicting labels at one pixel, corresponding to an occlusion, which are solved simply by assigning the label of the closest region in the conflict (these assignments are done implicitly by the order of processing the regions from farthest to closest in the Steps 2 to 7 of Algorithm 1). Also, there will be situations of disocclusions, where a pixel of A_{ij} will not receive

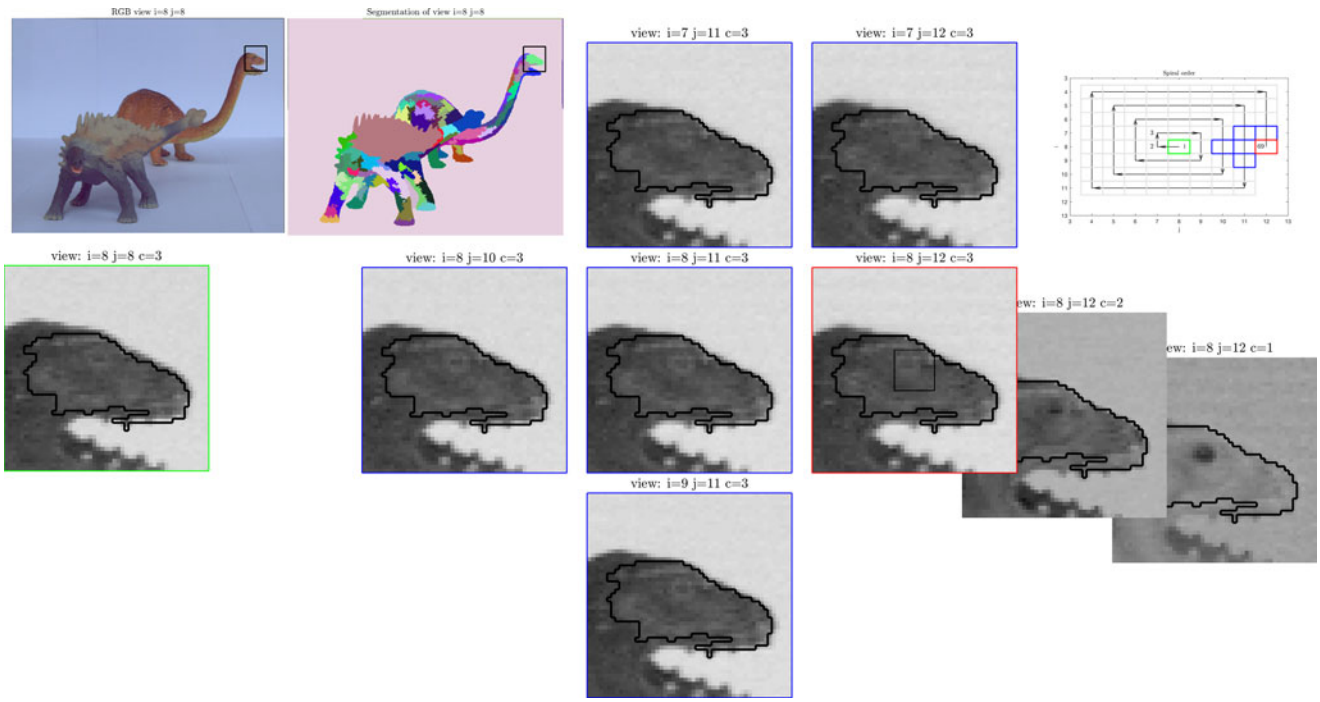


Fig. 2. The proposed method is able to remove redundancy of segmented regions by prediction due to the great similarity between neighboring views. One can note that the vertical and horizontal displacements of a region between two neighbor views are very small in the case of pictures taken with handheld plenoptic cameras. Here, we illustrate this observation by showing in the upper left corner one center view of the test images and its segmentation into regions in pseudo color. The blue color component of the highlighted area is shown in the green bordered image titled *view* $i = 8$ $j = 8$ $c = 3$. When moving along the spiral (see the upper right corner), at index $\zeta = 69$ (i.e. at $i = 8, j = 12$) the same, slightly shifted, area is shown in the red bordered image titled $i = 8$ $j = 12$ $c = 3$. The neighboring views that are already available to use in prediction and that belong to the set $\Gamma_{8,12}$ are shown bordered blue in their corresponding locations in the spiral. The color components that are already sent are also shown ($i = 8$ $j = 12$ $c = 2$ and $i = 8$ $j = 12$ $c = 1$). The contour of one region is highlighted in view $i = 8$ $j = 12$ $c = 3$ in black after region propagation and the very same contour is overlaid on the neighboring views (blue) and on the center view (green). It can be seen, that it fits to the neighboring views within one pixel but is clearly shifted compared to the center view. The area marked with a black square in *view* $i = 8$ $j = 12$ $c = 3$ corresponds to the patch in Fig. 3, where the structure of the linear model for sparse predictor dedicated to green color for this particular region is shown.

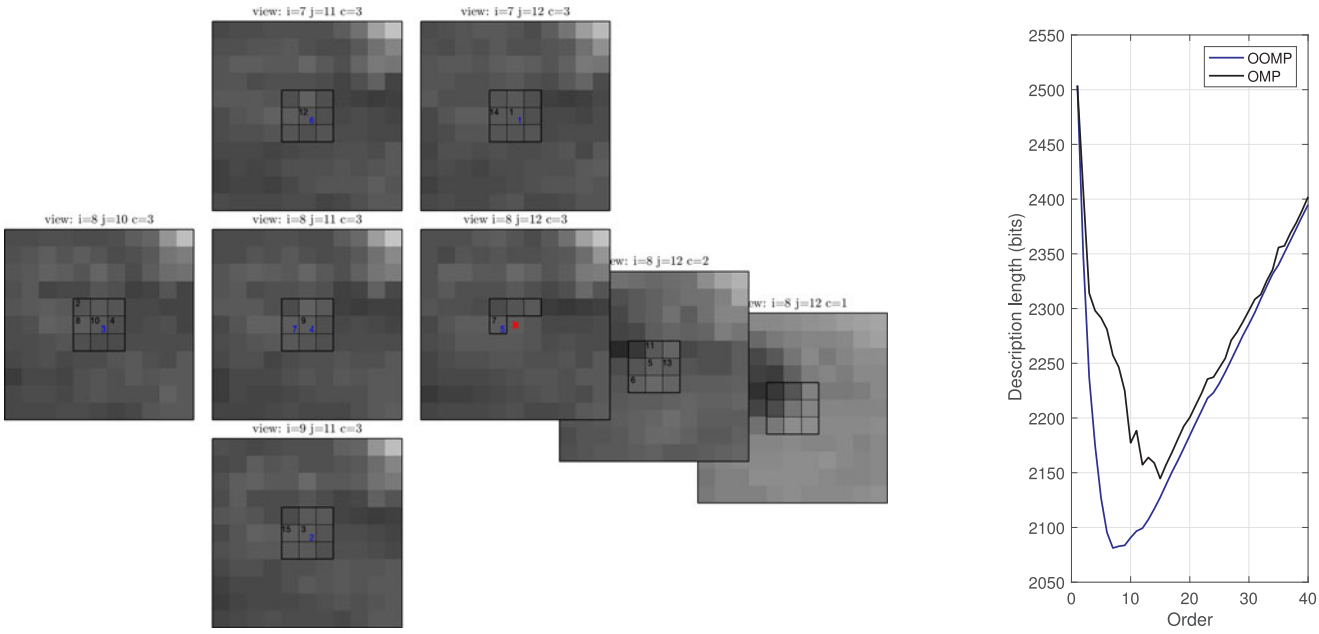
any label of a region (the pixels in the set Φ at Step 8). We use a simple method known as *neighbor's disparity assignment* [13] where the missing pixels are allocated to the nearby region (with respect to the coordinates x and y) located at the farthest depth, as described in steps 8–14 in Algorithm 1. An example showing the regions in a partition and the results of partition propagation can be seen in Figs. 2 and 4.

C. Encoding Scheme

The general scheme of the encoder is presented in Algorithm 2. The initially encoded elements are: the central view $A_{i_c j_c}$, which is encoded losslessly by JPEG 2000; and the central partition label image $Z_{i_c j_c}$, which has labels in the range $0, \ldots, L-1$ and is encoded by the lossless compressor CERV [14]. The impact of sending $Z_{i_c j_c}$ is small on the total code length (as will be discussed in Section V, the lossless code length needed is about 0.01% of the total code length, see Table III) so no major benefit could be obtained by using a lossy depth compression method such as [15]. Then the displacements for propagating the central partition to each side view is transmitted and the Algorithm 1 is used to construct the partition $\mathcal{P}_{ij}$ for every view. After the initial stage, the algorithm proceeds to encode all side views in the spiral order

Algorithm 1: Partition propagation from central view (i_c, j_c) to side view (i, j).

1: *Input:* The center view partition $\mathcal{P}_{i_c j_c} = \{\Omega^0_{i_c j_c}, \ldots, \Omega^{L-1}_{i_c j_c}\}$; disparities δ^k_{ij} and μ^k_{ij}.
2: **for** $k = L-1, \ldots, 0$ **do**
3: **for all** $(x, y) \in \Omega^k_{i_c, j_c}$ **do**
4: Set $x' = x + \delta^k_{ij}$ and $y' = y + \mu^k_{ij}$.
5: Insert (x', y') into Ω^k_{ij}.
6: **end for**
7: **end for**
8: Fill missing pixels $\Phi = \Omega \setminus \{\Omega^0_{ij}, \ldots, \Omega^{L-1}_{ij}\}$ to form partition:
9: **for** $k = L-1, \ldots, 0$ **do**
10: Dilate the set Ω^k_{ij} with a disk of radius r to obtain region Φ_1.
11: $\Phi_2 = \Phi \cap \Phi_1$ (pick from dilated region only missing pixels).
12: $\Omega^k_{ij} = \Omega^k_{ij} \cup \Phi_2$ (allocate to set Ω^k_{ij}).
13: $\Phi = \Phi \setminus \Phi_2$ (discard Φ_2 from missing pixels).
14: **end for**
15: *Output:* $\mathcal{P}_{ij} = \{\Omega^0_{ij}, \ldots, \Omega^{L-1}_{ij}\}$.



(a) (b)

Fig. 3. Illustration of sparse predictor design for a region of view $A_{ijc} = A_{8,12,3}$. An (11×11)-square where all pixels belong to the same region is shown in subfigures (a) and (b). (a) Illustration of OOMP template chosen by MDL. The pixel to be predicted, marked as a red $\times$ in "view $i = 8$ $j = 12$ $c = 3$," is set as the tth element of the vector $\mathbf{y}$. Its causal neighbors inside the black border regions are elements of the tth row of candidate regressors' matrix $\mathbf{D}$, together with all pixels inside the black contours in all shown views which are also candidate regressors. OOMP chooses sequentially to include in prediction template the pixels in the order marked by the overlaid blue numbers. Here MDL chooses to stop after selecting the first $k^* = 7$ regressors. OMP template chosen by MDL is denoted with black numbers. The optimal number of selected regressors is $k^* = 15$. Hence, OMP predictor is about twice slower than the OOMP predictor, which signifies about twice slower decoding speed. (b) The evolution of the MDL criterion for the whole region for the algorithms OOMP and OMP. OOMP chooses more effectively the regressor pixels for decreasing the MDL at the highest rate. (a) The available regressors and selected ones by OMP and OOMP. (b) Description length for the region.

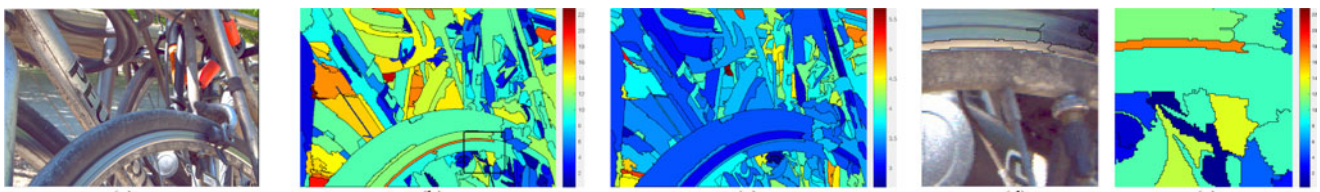


Fig. 4. (a) shows a color view of the dataset image *Bikes*. The cropped area marked with a black frame is zoomed-in in panel (d) and (e). (b) shows the C-segmentation of the side view, based on mean shift and partition propagation. The color in each region visualizes the value of self-to-cross ratio $\rho(\mathbf{\Theta})$ introduced in (11), in percents (see the color-to-value bar in the right of the panel). It can be seen that regions with uniform color, or regions with high displacement (they are either in front or far back) have high self-cross ratio, i.e., the causal mask for these regions is more important than the neighboring views. Regions with texture or low displacement obtain a low self-cross ratio. (c) shows the final compression performance as bits per pixel (per color component) value, represented by the color of the region (see the color-to-value bar in the right of the panel). Most difficult to compress areas are those far away, where the relative displacement between views is largest. (d) shows a zoom of (a) and overlays in black line the contours of regions after the partition propagation. (e) is a zoom in the corresponding area of (b) and it reveals the high self-cross ratio over almost uniform color regions. (a) The tenth side view, $(9, 6)$, of Bikes. (b) C-segmentation. Self-cross ratio shown as color. (c) C-segmentation. Code length (bit per pixel) as color. (d) Crop of (a). (e) Crop of (b).

presented in Fig. 1, by performing the loop at the Steps 6–11 of Algorithm 2, over all views and all regions and finally over all color components. Each side view A_{ij} is encoded using the side views that were already encoded as illustrated in Fig. 2 and described in the following. We restrict the set of views to be used when encoding the view A_{ij} to only the five closest side views, having indices (i', j') for which $|i - i'| + |j - j'|$ gets the smallest five values. In Fig. 2 the set of views for prediction is $\Gamma_{8,12} = \{(8, 11), (7, 12), (7, 11), (9, 11), (8, 10)\}$. For each view (i, j) the pixels in the region Ω_{ij}^k are encoded by predictive coding, using a sparse linear predictor, specifically designed for the region Ω_{ij}^k and sent along to decoder as side information. The linear prediction model uses as regressors the pixels from the side views already encoded and from the current side view. The detailed model description is presented in Section II-D.

D. Linear Regression Model Setting for One Region of One View

In this section we describe the structure of the regression problems that are solved in Step 5 of Algorithm 2. Suppose that the decoder has already received and decoded completely the views $A_{8,8}, \ldots, A_{7,12}$ along the spiral from Fig. 1, and the encoder needs to encode the view $A_{8,12}$. The situation is depicted in

Algorithm 2: Encoder.

Input: The 5D light-field L_F and the partition of the center view $\mathcal{P}_{i_c j_c} = \{\Omega^0_{i_c j_c}, \ldots, \Omega^{L-1}_{i_c j_c}\}$.
Find region displacements δ^k_{ij} and μ^k_{ij} for all regions Ω^k_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, N$, $k = 0, \ldots, L-1$ according to (1). Write the displacements into the bitstream using $2 \times 5 \times L \times N^2$ bits.
Obtain partitions for side views using Algorithm 1.
Encode center view $A_{i_c j_c}$ using lossless mode of JPEG 2000. Encode center partition label image $Z_{i_c j_c}$ with CERV.
Advance in sequential order along the spiral and encode each view exploiting the already encoded views, as follows:
for all i, j, c, k **do**
  (for each side view (i, j), color c, and region Ω^k_{ij})
  Design using the algorithm Algorithm 4 a sparse predictor $\mathbf{\Theta}^k_{ijc}$, by selecting the regression matrix $\mathbf{D}$ and the desired vector $\mathbf{y}$, as described in Section II-D.
  Encode each sparse predictor, $\mathbf{\Theta}^k_{ijc}$ as described in Section II-E.
  Encode the prediction residuals $\boldsymbol{\varepsilon} = \mathbf{y} - \lfloor \mathbf{D}\mathbf{\Theta}^k_{ijc} \rceil$ for each color c, and region Ω^k_{ij} as described in Section II-F.
end for

Algorithm 3: Decoder.

Input: Encoded displacements, encoded predictors, encoded residuals.
Read region displacements δ^k_{ij} and μ^k_{ij} for all $i = 1, \ldots, N$, $j = 1, \ldots, N$, $k = 0, \ldots, L-1$.
Decode the center view $A_{i_c j_c}$ using the lossless mode of JPEG 2000. Decode the center partition $\mathcal{P}_{i_c j_c} = \{\Omega^0_{i_c j_c}, \ldots, \Omega^{L-1}_{i_c j_c}\}$ with CERV.
Obtain partitions for side views using Algorithm 1.
Advance in sequential order along the spiral and decode each view as follows:
for all i, j, c, k **do**
  (for each side view (i, j), color c, and region Ω^k_{ij})
  Decode each sparse predictor, $\mathbf{\Theta}^k_{ijc}$.
  for all ℓ **do**
    Form the ℓth row $\mathbf{d}^T_\ell$ of the regression matrix $\mathbf{D}$ from the already decoded side views as described in Section II-D.
    Decode the prediction residual ε_ℓ (the ℓth residual for color c, and region Ω^k_{ij}).
    Obtain the reconstructed value $y_\ell = \lfloor \mathbf{d}^T_\ell \mathbf{\Theta}^k_{ijc} \rceil + \varepsilon_\ell$, which represents the value of color c for the ℓth pixel from region Ω^k_{ij}.
  end for
end for

Fig. 2. The regions $\Omega^0_{8,8}, \ldots, \Omega^{L-1}_{8,8}$ of the central view are propagated using the displacements $(\delta^0_{8,12}, \mu^0_{8,12}), \ldots (\delta^{L-1}_{8,12}, \mu^{L-1}_{8,12})$ respectively, to get the corresponding regions $\Omega^0_{8,12}, \ldots, \Omega^{L-1}_{8,12}$. The contour of one region is highlighted by black border in Fig. 2. The view $A_{8,12}$ is encoded region-by-region, selecting the regions in their row-wise scanning order. Next, we describe the encoding of a generic region, $\Omega^k_{8,12}$. The color components at all pixels in the region are encoded sequentially: first the red component ($c = 0$) is transmitted for all pixels in $\Omega^k_{8,12}$, then the green component $c = 1$ and last, the blue component $c = 2$, e.g., in Fig. 2, we encode color component $c = 2$, hence the first two color components have already been sent. A separate linear regression model is built for every color component, $c \in \{0, 1, 2\}$. Let the 3×3 square centered at (x, y) be $\mathcal{N}_{3\times3}(x, y) = \{(x + x_1, y + y_1) | (x_1, y_1) \in \{-1, 0, 1\}^2\}$. The model expresses the value of $A_{8,12}(x, y, c)$ at a pixel $(x, y) \in \Omega^k_{8,12}$ as a linear combination of the color values at all pixels $(x', y') \in \mathcal{N}_{3\times3}(x, y)$ in the view $A_{8,12}$ that are already transmitted to the decoder, and of the color component c at all pixels $(x', y') \in \mathcal{N}_{3\times3}(x, y)$ in the selected neighbor views $A_{i',j'}$ with $(i', j') \in \Gamma_{ij}$. A highlighted square of pixels of the encoded view in Fig. 2 is enlarged in Fig. 3 where the model is seen such that (x, y) is denoted by a red $\times$ and the pixels surrounded by black border belong to the model. The linear model is expressed in a matrix form

$$\mathbf{y} = \mathbf{D}\mathbf{\Theta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{D} \in \mathbb{R}^{n \times m}$, $\mathbf{\Theta} \in \mathbb{R}^m$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$. Every value $A_{8,12}(x, y, c)$ with $(x, y) \in \Omega^k_{8,12}$ appears as an entry in the vector $\mathbf{y}$, hence the size of the vector $\mathbf{y}$ is equal to the cardinality of the region $\Omega^k_{8,12}$ (in our experiments the cardinality $|\Omega^k_{8,12}|$ starts from few hundreds up to several tens of thousands). The regression matrix $\mathbf{D}$ contains the elements in the linear combinations, explicitly given below, and the linear regression coefficients are collected in the vector of unknowns, $\mathbf{\Theta}$. The elements of the $c-$th color plane in the region $(x, y) \in \Omega^k_{8,12}$ are inserted in $\mathbf{y}$ according to the row-wise scanning order. All the elements are inserted in the regression matrix $\mathbf{D}$ because they are possibly relevant to the values of $\mathbf{y}$. However, their relevance and their being part of the final linear model will be decided by the sparse design algorithm. We give explicitly the structure of the linear system $\mathbf{y} = \mathbf{D}\mathbf{\Theta} + \boldsymbol{\varepsilon}$ by describing its generic row r. The rth element in $\mathbf{y}$ is $A_{i,j}(x, y, c)$ for some $(x, y) \in \Omega^k_{i,j}$, and it is what we aim to predict by a linear combination of already known pixel color values. The rth row in $\mathbf{D}$ consists of the elements that are considered relevant to $A_{i,j}(x, y, c)$, which are listed in the groups below:

1) the *self*-prediction group contains the color c value at the west, north-west, north, and north-east neighbors, i.e., include in the row $\mathbf{D}$ the four elements, $A_{ij}(x, y - 1, c)$, $A_{ij}(x - 1, y - 1, c)$, $A_{ij}(x - 1, y, c)$, and $A_{ij}(x - 1, y + 1, c)$. They form the *causal* set denoted $\mathcal{N}^c(x, y)$. The set $\mathcal{N}^c(x, y)$ is causal in the row-wise scanning of the region Ω^k_{ij}.
2) the *cross*-prediction, across the other color-planes in the same view, whenever the color-component $c - 1$ or

$c-2$ exists, i.e., include in the row $\mathbf{D}$ the nine elements $A_{ij}(x',y',c-1)$ if $c \geq 1$ and additionally $A_{ij}(x',y',c-2)$ if $c=2$, with $(x',y') \in \mathcal{N}_{3\times 3}(x,y)$.

3) the *cross*-prediction across the other views, only at the color component c, i.e., for every view $(i',j') \in \Gamma_{ij}$ include in the row $\mathbf{D}$ the nine elements $A_{i'j'}(x',y',c)$, with $(x',y') \in \mathcal{N}_{3\times 3}(x,y)$.
3) the first element in each row is a constant, 1, for modeling the possible bias of the model.

After building the matrix $\mathbf{D}$ as above, one adjustment step is done, with the goal to constrain all elements in the regression matrix to be only elements of the region Ω_{ij}^k, as in [16]. When a pixel $(x',y') \in \mathcal{N}_{3\times 3}(x,y)$ does not belong to the region Ω_{ij}^k the color value at that pixel is replaced with the corresponding color value of the nearest pixel in the neighborhood that belongs to the region (in the same way as in [16]). Similarly, if $(x',y') \in \mathcal{N}^c(x,y)$ does not belong to the region Ω_{ij}^k it is replaced by the closest from the causal template $\mathcal{N}^c(x,y)$. If this is not possible, i.e., if no pixel in the causal template $\mathcal{N}^c(x,y)$ belongs to the region Ω_{ij}^k, the encoder (and the decoder as well) will not consider (x,y) to be predicted by the specific predictor for color c and region Ω_{ij}^k, but instead will be treated by a general predictor, one for each view, designed especially for these "border" pixels, that uses only the cross-terms of the prediction model listed above. For views where already five previous views were encoded, the matrix $\mathbf{D}$ will have the following number, m, of columns (possible regressors): $m=50$ for the red component ($c=0$); $m=59$ for green component ($c=1$); and $m=68$ for blue component ($c=2$). The task of the sparse predictor is to select only the relevant elements in the regression model, by selecting only those columns that contribute to increasing the accuracy of prediction, otherwise their presence in the model may be detrimental, because they incur a bit-cost for being encoded along in the bitstream. Especially for small regions, this cost may be relatively high part of the overall residual-plus-coefficients cost. We denote by $\mathbf{\Theta}$ a sparse predictor, where only the elements $\mathbf{\Theta}_i$ with indices i in the support $\mathcal{I} \in \{1,\ldots,m\}$ are nonzero. The nonzero elements from the vector are denoted $\boldsymbol{\theta} = \mathbf{\Theta}_{\mathcal{I}}$ and the model prediction can be written as $\hat{\mathbf{y}} = \mathbf{D}\mathbf{\Theta} = \mathbf{D}_{\mathcal{I}}\boldsymbol{\theta}$, where $\mathbf{D}_{\mathcal{I}}$ is selecting only the columns with the indices belonging to $\mathcal{I}$ from the matrix $\mathbf{D}$. The matrix $\mathbf{D}$ in sparse modelling is sometimes called a dictionary, especially when it is determined using training data, different to the data where the regression model is finally used. In here however, we construct $\mathbf{D}$ specifically for the data to be encoded, and we have one such $\mathbf{D}$ for each of the regions, each of the views, and each of the color components, resulting in e.g., $|\Omega_{ij}^k| \times N^2 \times 3 = 70000$ different regression matrices, each specific to the data where it has to be used. The regression matrix $\mathbf{D}$ formed in this way includes columns with high mutual coherence, which is known to be a challenging situation for sparse recovery algorithms. We introduce the investigated algorithms in Section III and show their performance in Section V. We do not perform normalizing and centering of the regression matrix $\mathbf{D}$, which will need sending additional information about means and variances of the columns since the decoder is unaware in the beginning of the full matrix $\boldsymbol{D}$ (which is recovered row by row in the process of decoding $\mathbf{y}$) and cannot compute all means and variances from columns. However, we observed that the lack of centering and normalizing does not deteriorate the performance of sparse recovery algorithms presented in Section III.

E. Encoding the Predictor Coefficients

We consider a sparse predictor $\mathbf{\Theta}$ with m real valued elements, having the support $\mathcal{I}$ of k nonzero elements $\mathbf{\Theta}_{\mathcal{I}} = \boldsymbol{\theta}$. In order to encode $\mathbf{\Theta}$, first the nonzero elements are rounded to keep only N_b bits in the fractional part, resulting in the quantized vector $\overline{\boldsymbol{\theta}}$ with elements:

$$\overline{\theta}_i = \lfloor \theta_i 2^{N_b} \rceil / 2^{N_b};\ \ i = 1,\ldots,k \tag{2}$$

which are transmitted to the decoder in the form of the integer values $m_i = \lfloor \theta_i 2^{N_b} \rceil$, $i=1,\ldots,k$. If by quantization some of the coefficients of $\boldsymbol{\theta}$ become zero, that will be treated by changing the definition of the support $\mathcal{I}$ to point only to non-zero coefficients and setting k to the size of the new $\mathcal{I}$.

1) First the value of $k \in \{1,\ldots,K\}$ is encoded, using $\lceil \log_2 K \rceil$ bits. The signs of the non-zero coefficients are then sent, without coding, as a k-long binary string.
2) The indices of the support $\mathcal{I}$ are transmitted as follows. A binary vector $\boldsymbol{\gamma} = [\gamma_1,\ldots,\gamma_m]$ is created having ones at the indices from $\mathcal{I}$. This binary vector is encoded using $\log_2 \binom{m}{k}$ bits, which amounts to assuming that all m-long binary sequences of weight k (i.e., having k symbols one) are equally likely, having probability $\frac{1}{\binom{m}{k}}$ each. One way to implement this code is to enumerate all such strings (e.g. using the indexing scheme presented in [17]), and then to encode, using arithmetic coding, the resulting global index of the particular $\boldsymbol{\gamma}$ (which is an integer between 1 and $\binom{m}{k}$). However, we implement this in a simpler way, by sequentially encoding the binary symbols of $\boldsymbol{\gamma} = [\gamma_1,\ldots,\gamma_m]$, and hence avoiding the use of global indices, which may be too large. At the first element, γ_1, the probability distribution is defined by $P(\gamma_1 = 1) = \binom{m}{k}$ and then for the ith element, $P(\gamma_i = 1) = \frac{\sum_{j=i}^{m} \gamma_j}{m-i+1}$, where $i = 1,\ldots,m$. In fact at most $m-1$ binary symbols need to be encoded; the encoding process will stop whenever $P(\gamma_j = 1) = 1$ or $P(\gamma_j = 1) = 0$, which may already happen at a j less than $m-1$. One can easily show that overall this encoding process results in the assignment $P(\boldsymbol{\gamma}) = \frac{1}{\binom{m}{k}}$.
3) The absolute values are encoded using a Golomb Rice code of parameter p, where encoding a positive integer m is done by encoding in unary the quotient $\lfloor \frac{m}{2^p} \rfloor$ and sending the p bits of the remainder unencoded. The unary part of the code is formed of $\lfloor \frac{|m_i|}{2^p} \rfloor$ bits followed by a 0 separator, requiring in total $k + \sum_{i=1}^{k} \lfloor \frac{|m_i|}{2^p} \rfloor = k + \sum_{i=1}^{k} \lfloor \frac{\lfloor |\theta_i| 2^{N_b} \rceil}{2^p} \rfloor$ bits, while the unencoded remainders are needing kp bits.

Algorithm 4: Greedy sparse predictor design algorithm with MDL structure selection, having a common structure for OOMP and OMP algorithms. OOMP solves a full LS problem in the inner loop at step 7, while OMP basically performs the scalar product between $\mathbf{r}_{k-1}$ and $\mathbf{D}_i$ in the inner loop at step 9 (all $\mathbf{D}_i^T\mathbf{D}_i$ are precomputed outside the loop).

1: *Input:* $\mathbf{y}, \mathbf{D}$
2: *Initialize:* $\mathcal{I}_0 = \emptyset$; $\mathcal{A}_0 = \{1, \ldots, m\}$
// Outer loop
3: **for** $k = 1$ to K **do**
4: *// Inner loop*
5: **for** $i \in \mathcal{A}_{k-1}$ **do**
6: **if** Algorithm $=$ OOMP **then**
7: $J_i = \min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{D}_{\mathcal{I}_{k-1}\cup i}\boldsymbol{\theta}\|^2$
8: **else if** Algorithm $=$ OMP **then**
9: $J_i = \|\mathbf{r}_{k-1}\|^2 - \frac{(\mathbf{r}_{k-1}^T\mathbf{D}_i)^2}{\mathbf{D}_i^T\mathbf{D}_i}$
10: **end if**
11: **end for**
12: $i^* = \arg\min_i J_i$
13: $\mathcal{I}_k = \mathcal{I}_{k-1} \cup i^*$
14: $\mathcal{A}_k = \mathcal{A}_{k-1} \setminus i^*$
15: $\boldsymbol{\theta}_k = \min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{D}_{\mathcal{I}_k}\boldsymbol{\theta}\|^2$
16: $\mathbf{r}_k = \mathbf{y} - \mathbf{D}_{\mathcal{I}_k}\boldsymbol{\theta}_k$
17: $\overline{\boldsymbol{\theta}}_k = Q(\boldsymbol{\theta}_k)$ (see (2))
18: $\mathrm{MDL}_k = \mathrm{MDL}(\mathcal{I}_k, \overline{\boldsymbol{\theta}}_k)$ (see (8))
19: **end for**
20: $k^* = \arg\min_k \mathrm{MDL}_k$
21: *Output:* $k^*, \mathcal{I}_k, \overline{\boldsymbol{\theta}}_k$

Summarizing, the overall code length for encoding the coefficients is

$$\mathcal{L}_{\boldsymbol{\theta}} = \sum_{i=1}^{k}\left\lfloor \frac{\lfloor |\theta_i| 2^{N_b}\rceil}{2^p}\right\rfloor + \log_2\binom{m}{k} + k(p+2) + \lceil \log_2 K\rceil. \tag{3}$$

The optimal Golomb-Rice parameter p^* is found evaluating (3) over the predefined set $p \in \{N_b - i_1, \ldots, N_b + i_2\}$ and picking p^* that minimizes $\mathcal{L}_{\boldsymbol{\theta}}$. In our experiments $i_1 = 2, i_2 = 5$ and the value of p^* is hence encoded using 3 bits, which is a constant value and is neglected in the expression of the code length for brevity, resulting in the optimal code length for coefficients

$$\mathcal{L}^*_{\boldsymbol{\theta}} = \sum_{i=1}^{k}\left\lfloor \frac{\lfloor |\theta_i| 2^{N_b}\rceil}{2^{p^*}}\right\rfloor + \log_2\binom{m}{k} + k(p^*+2) + \lceil \log_2 K\rceil \tag{4}$$

The expression (4) is evaluated quickly, at every stage of the outer loop in the Algorithm 4 and in the Step 5 of Algorithm 5, as part of the MDL criterion.

F. Encoding the Prediction Errors

The encoder is using for prediction the predictor coefficients that are available to the decoder as well, i.e., the predictor coefficients $\overline{\boldsymbol{\theta}}$ from (2), and the residual vector

Algorithm 5: LARS-LASSO sparse predictor design algorithm with MDL structure selection.

1: *Input:* $\mathbf{y}, \mathbf{D}$
2: Run LARS algorithm and obtain the sequence $\{\mathcal{I}_k, \boldsymbol{\theta}_k\}_{k=1,\ldots K}$.
3: **for** $k = 1$ to K **do**
4: $\overline{\boldsymbol{\theta}}_k = Q(\boldsymbol{\theta}_k)$ (see Eq (2))
5: $\mathrm{MDL}_k = \mathrm{MDL}(\mathcal{I}_k, \overline{\boldsymbol{\theta}}_k)$ (see Eq (8))
6: **end for**
7: $k^* = \arg\min_k \mathrm{MDL}_k$
8: *Output:* $k^*, \mathcal{I}_k, \overline{\boldsymbol{\theta}}_k$

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{D}_{\mathcal{I}}\overline{\boldsymbol{\theta}} \tag{5}$$

having n elements $\varepsilon_1, \ldots, \varepsilon_n$ needs to be encoded for lossless reconstruction. The final encoding of the prediction errors (called for short also residuals) in the bitstream is done using a rather involved context coding machinery, typical for most lossless image compression applications. The contexts are defined, in a similar manner to [18], [19], according to the quantized values of the differences between neighbor pixels in the causal neighborhood of the current pixel (see [16] for the exact definitions, used also here). At each context, the minimum and maximum of the value to be encoded are denoted $s_{\min}, s_{\max}$ respectively, and are encoded as side information. A histogram is adaptively maintained at each context, where the current counts $N(s_{\min}), \ldots, N(s_{\max})$ for all symbols are used for defining the probability $p(\varepsilon_t = s) = \frac{N(s)+1}{\sum_{i=s_{\min}}^{s_{\max}}(N(i)+1)}$ for encoding the current symbol $\varepsilon_t = s$, and then $N(s)$ is incremented. This nonparametric way of encoding ensures a robust performance, irrespective of the degree of matching of the real distribution of the residuals to any assumed model. Although in practice the above context coding is used for creating the bitstream, during the sparse design of the linear predictor one needs to make assumptions about the ideal distribution of the residuals. We assume a Gaussian distribution for the residuals, and use it for its convenience in allowing recursive in order least squares equations, achieving a fast predictor design. Similar to the MDL scheme used in [20] we assume the distribution of the residuals given by

$$p(\boldsymbol{\varepsilon}|\hat{\sigma}^2) = \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}} e^{-\sum_{t=1}^{n}\varepsilon_t^2/(2\hat{\sigma}^2)},$$

where $\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2$. The optimal code length for residuals equals the negative maximized log-likelihood

$$\begin{aligned}\mathcal{L}_{\varepsilon}(\boldsymbol{\varepsilon}) &= -\log_2 \frac{1}{(2\pi\hat{\sigma}^2)^{n/2}} e^{-\sum_{t=1}^{n}\varepsilon_t^2/(2\hat{\sigma}^2)} \\ &= \frac{n}{2}\log_2(2\pi) + \frac{n}{2}\log_2\hat{\sigma}^2 + \frac{n}{2}\log_2 e \\ &= \frac{n}{2}\log_2\sum_{t=1}^{n}\varepsilon_t^2 + \frac{n}{2}\log_2(2\pi e/n) \\ &= \frac{n}{2}\log_2\|\mathbf{y} - \mathbf{D}_{\mathcal{I}}\overline{\boldsymbol{\theta}}\|^2 + \frac{n}{2}\log_2(2\pi e/n). \end{aligned} \tag{6}$$

This description length is achievable in practice by using arithmetic coding, so we call it an *implementable* description length, but in the final stage of constructing the bitstream we are using the usually more efficient code length given by the context based coding described above.

III. Sparse Modeling

A. The Implementable MDL for Model Structure Selection

One instance of the sparse design problem is defined by the $n \times m$-matrix $\mathbf{D}$ and the n-vector $\mathbf{y}$ and the task is to find a (sub)optimal sparse predictor with support $\mathcal{I}_{k^*} \subset \{1, \ldots, m\}$ and the nonzero predictor parameters $\overline{\boldsymbol{\theta}} \in \mathcal{Q}_{N_b}^{k^*} \subset \mathbb{R}^{k^*}$, where $\mathcal{Q}_{N_b} = \{i/2^{N_b} | i \in \mathbb{Z}\}$ is the subset of real numbers whose fractional part is represented with N_b bits. The optimality criterion is the overall description length, MDL_{k^*}, which includes the code lengths for the prediction coefficients $\overline{\boldsymbol{\theta}}$, the prediction support $\mathcal{I}_{k^*}$ and prediction residuals

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{D}_{\mathcal{I}_{k^*}} \overline{\boldsymbol{\theta}}. \tag{7}$$

According to the considered encoding scheme, the implementable minimum description length can be computed using (4) and (6) as follows:

$$\begin{aligned} \mathrm{MDL}(\mathcal{I}, \overline{\boldsymbol{\theta}}) &= \mathcal{L}_{\varepsilon}(\boldsymbol{\varepsilon}) + \mathcal{L}_{\boldsymbol{\theta}}^{*} = \frac{n}{2} \log_2 \|\mathbf{y} - \mathbf{D}_{\mathcal{I}} \overline{\boldsymbol{\theta}}\|^2 \\ &+ \frac{n}{2} \log_2 (2\pi e/n) \\ &+ \sum_{i=1}^{m} \left\lfloor \frac{\lfloor |\theta_i| 2^{N_b} \rceil}{2^{p^*}} \right\rfloor + \log_2 \binom{m}{k} + k(p^* + 2) + \lceil \log_2 K \rceil, \end{aligned} \tag{8}$$

where k is the cardinality of the support $\mathcal{I} \subset \{1, \ldots, m\}$, and p^* depends on $\overline{\boldsymbol{\theta}}$ as presented in Sec. II-E.

Minimizing the description length is clearly the main goal in our application, and additionally, a sparse solution having a small support size $k^* = \|\overline{\boldsymbol{\Theta}}\|_0$ is desirable since it will lead to a fast decoding time. Using typical sparse designs based on greedy algorithms is convenient for providing sparse solutions $\overline{\boldsymbol{\Theta}}_k$ minimizing $\|\mathbf{y} - \mathbf{D}\overline{\boldsymbol{\Theta}}\|^2$ for each candidate support size $k = \|\overline{\boldsymbol{\Theta}}_k\|_0$ starting from 1 up to an order K fixed by the user, and hence the minimizing MDL_{k^*} will be found by inspecting the evaluated criteria at the sparse solutions, $\mathrm{MDL}_1 \ldots, \mathrm{MDL}_K$.

B. The Generic Structure of the Algorithm for Sparse Design and Model Structure Selection

We need to find the optimal sparse predictors for all regions in all side views, which in typical cases requires tens of thousands of predictor designs for one plenoptic image (in experiments we have 224 side views and in each view there are hundreds of regions, hence tens of thousands of predictor designs are not uncommon for one image). For this reason we consider here only sparse design methods which are very fast for the low dimensional models encountered in this application, where typically n is in the thousands while m is in the tens. We investigated the performance of three algorithms: orthogonal matching pursuit (OMP), optimized orthogonal matching pursuit (OOMP), and least angle regressor (LARS) that were found computationally feasible in the context of our problem. The first two belong to the greedy class of algorithms and have similar algorithmic structure, where the solutions are constructed recursively in the sparsity k of the predictor, i.e., in the number of nonzero elements of the predictor. The solutions are constructed for all $k = 1, 2, \ldots, K$ where K is the maximum number of regressors variables allowed by the user in the model, out of all m existing regressors. The optimal value of k is then selected according to the minimum description length criterion, which in this lossless compression application is the only genuinely relevant criterion. In a greedy algorithm the set $\{1, \ldots, m\}$ of all regressor indices is split at each stage k into the chosen regressors set, $\mathcal{I}_k$ and available regressors $\mathcal{A}_k$. The algorithm progresses from stage $k-1$ by enlarging the predictor support $\mathcal{I}_{k-1}$ (which was found the best at sparsity $k-1$) with a new regressor, to obtain the support $\mathcal{I}_k$, having cardinality k. Algorithm 4 shows the common structure of the sparse predictor design, where the inner loop makes the difference between the more complex OOMP algorithm or the faster OMP algorithm. After the inner loop is terminated, the best index i^* for extending the support is found and used to update the support $\mathcal{I}_k$ and the set $\mathcal{A}_k$ of available indices for next outer loop stage. For the selected support $\mathcal{I}_k$ the LS solution $\boldsymbol{\theta}_k$ is computed at line 15, and the residual $\mathbf{r}_k$ corresponding to it is computed at line 16 (that is strictly needed only for the OMP algorithm). The LS parameters $\boldsymbol{\theta}_k$ are then rounded to keep only N_b bits in their fractional part, by the uniform quantizer described in (2) at Appendix II-E. Finally the MDL_k is evaluated. After terminating the outer loop, the best k^* can be picked by inspecting $\mathrm{MDL}_1, \ldots, \mathrm{MDL}_K$. The algorithms OOMP, OMP and LARS are presented briefly in the next subsections.

C. The Algorithm OOMP

The sparse design algorithm OOMP, also known under various names, e.g., LS-OMP, Greedy LS, Fast OLS, and order-recursive forward selection, basically solves the problem of forward selection in least squares modeling, which appears in many application fields: econometry [21], control and system identification [22], optimization [23], sparse representations [24]. Despite the different guises of the problem, its mathematical structure is however the same, and the fast solutions presented in e.g. [21]–[24] are similar as main algorithmic steps, differing mainly in the trade-off between using more or less variables (and memory space) versus getting a faster or slower execution. In this paper we use the fast and space efficient algorithm Fast Orthogonal Least Squares (FOLS) from [25].

D. The Algorithm OMP

The second investigated algorithm is OMP, which evaluates scalar products at each inner iteration (line 9 in Algorithm 4), and can be seen as a sped-up version of OOMP. The new regressor index being selected is the one which gives the maximum in

$$i^* = \arg \max_{i \in \mathcal{A}_{k-1}} \frac{(\mathbf{r}_{k-1}^T \mathbf{D}_i)^2}{\mathbf{D}_i^T \mathbf{D}_i} \tag{9}$$

where the residual vector at stage $k-1$ is $\mathbf{r}_{k-1} = \mathbf{y} - \mathbf{D}_{\mathcal{I}_{k-1}} \boldsymbol{\theta}^*_{k-1}$.

E. The Algorithm LARS

Finally, we also consider the sparse design method LARS, which bridges two main classes of sparse design: the greedy algorithms and the convex relaxation algorithms, being a method able to provide solutions to the LASSO problem

$$\min_{\mathcal{I}\subset\{1,\ldots,m\}} \min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{D}_{\mathcal{I}}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|_1 \tag{10}$$

for all values of the regularization factor λ. The regularization path for LASSO is a sequence of predictor supports $\mathcal{I}_1, \ldots, \mathcal{I}_K$ having sparsity $k = |\mathcal{I}_k|$, where the regularization factor λ is iteratively set by the LARS algorithm so that the resulting sparsity is k. For details we refer to the original paper describing LARS [26] and the implementation of the algorithm from [27]. The program from [27] has the option of running the LASSO type of LARS algorithm, in which case the support $\mathcal{I}_k$ does not necessarily include $\mathcal{I}_{k-1}$ and does not necessarily have cardinality k; hence $\mathcal{I}_1, \ldots, \mathcal{I}_K$ are not nested. In Algorithm 5 are shown the steps necessary to include the LARS-LASSO algorithm in the structure for MDL predictor structure selection, where the condition of nested supports is not assumed.

IV. Partitioning Methods

A. Simple Partitioning Methods

In [12] the partition of the central view $\mathcal{P}_{i_c,j_c} = \{\Omega^0_{i_c,j_c}, \ldots, \Omega^{L-1}_{i_c,j_c}\}$ that needs to be input to Algorithm 2 was obtained simply by quantizing the disparity map. In this paper we consider more refined means of obtaining the partition. Any arbitrary segmentation of $A_{i_c j_c}$ can be used directly in our compression scheme, when denoting the segments $\{\Omega^0_{i_c j_c}, \ldots \Omega^{L-1}_{i_c j_c}\}$, ordered according to the average depth over the region, so that $\Omega^{k-1}_{i_c j_c}$ corresponds to a region that is closer in depth than the region $\Omega^k_{i_c j_c}$. This re-labeling can be done using an explicit estimated depth map or simply using the estimated disparities discussed as a part of the encoding method in Section II.

The selected *self* regressors in the sparse predictor (those that belong to the current view to be encoded, see Section II-D) are able to contribute with best interpolation of the color from the neighbor pixels, so that homogeneity of a region can be evaluated in terms of the similarity of the colors or texture of the region, and hence a color segmentation is a good candidate to be investigated. If only the self part would be present, a human subjective segmentation would be very good. However, the *cross* part of the predictor (which exploits the cross-correlation of the current pixel with regressors from neighbor views) is implementing an interpolation process based on the geometrically relevant pixels, according to the relative position of the lenslets and also depending on the distance of the object in the 3D scene, leading to ideal regions carved in terms of the (quantized) depth of the pixels. Hence an optimal partition is not necessarily a good human segmentation of the color image, but also involves the optimal carving of regions based on the depth values.

We quantify the relative importance between *self* regressors and *cross* regressors, as established by the optimal sparse design, by splitting the support $\mathcal{I}$ of the sparse predictor into two subsets: the self-support $\mathcal{I}_{\text{self}}$ which points to pixels in the causal prediction mask from own view, and the cross-support $\mathcal{I}_{\text{cross}}$ which points to pixels in the neighbor views and defining the self-to-cross ratio

$$\rho(\boldsymbol{\Theta}) = \frac{\sum_{i\in\mathcal{I}_{\text{self}}} |\Theta_i|}{\sum_{i\in\mathcal{I}_{\text{cross}}} |\Theta_i|}. \tag{11}$$

Our MDL criterion automatically selects the non-zero elements of the predictor for an optimal prediction, hence one can see $\rho(\boldsymbol{\Theta})$ as giving hints about which local homogeneity is more important: that based on color, when ρ is high, or that based on depth, when ρ is low. In Fig. 4 are shown the self-to-cross ratios over each region of a mean-shift segmentation of the color view; the values of ρ are represented acording to the vertical color-bar next to the image.

1) Disparity-Map Partitioning Method, D: When predicting the color value at a pixel (x, y) from view (i, j), one can see the optimal sparse predictor as an optimal interpolator of the values present in the prediction mask, which may belong to a neighbor view (i', j'), among others. In the case of a view (i, j) which is shifted by a fractional disparity with respect to the neighbor view (i', j'), the interpolator will adjust itself according to the disparity, and if the disparity is constant over a certain region, the same predictor will be optimal for the whole region. Hence segmentation based on disparity makes a good basis. We adopted this strategy earlier in [12] and now we show how it compares to other segmentation alternatives in Section V. The depth map for plenoptic images can be estimated with numerous methods, see e.g. [28]. In this paper we use the depth map provided by the camera manufacturer's software, namely Lytro Desktop. Given a pixel with disparity d we quantize it to $d_q = \lfloor N_d \frac{d-d_{\min}}{d_{\max}-d_{\min}} \rceil$ where N_d is the number of constant disparity levels and $d_{\max}$ and $d_{\min}$ correspond to the range of disparities. The final quantized disparity map emerges from the pixels that have the the same quantized disparity and belong to the same group of connected pixels. We denoted in results the depth-map partitioning method with suffix D. In practice, the regions obtained with quantized disparity maps lack in quality when it comes to object contours, due to the fact that light-field depth estimation methods, such as [29], rely on mixing the results of two approaches: first, the traditional stereo image matching method, and second, the so-called depth-from-defocus, where depth is deduced by refocusing on different depths, with precision limited by the angular resolution $N \times N$.

2) Color Partitioning Method, C: In order to have regions with more precise contours, we consider also color segmentations, offering good object contours but regions containing sometimes a wide range of disparities. The result when using as color segmentation the mean shift algorithm [30] are specified by the suffix C in the method name.

3) Region Merging Based on Depth Similarity, D − RM: In order to obtain constant disparity regions with precise image object contours, the color and depth information is combined by starting from a color oversegmentation and exploiting depth in

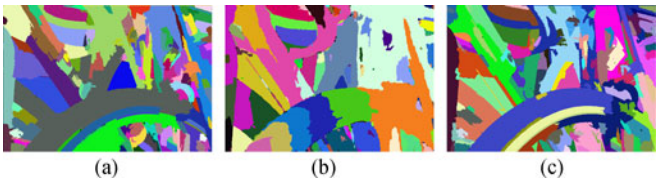


Fig. 5. The three different simple methods to obtain a partition introduced in Section IV-A. All partitions have about 130 regions. (a) Color segmentation, -C. (b) Constant disparity segmentation, -D. (c) Constant disparity with region merging, -D-RM.

a region merging process where we merge neighboring regions that are the closest in depth until a certain number of regions is obtained. For regions $\Omega_{ij}^{k_1}$ and $\Omega_{ij}^{k_2}$, we define a merging criterion as

$$\mathcal{J}_d = \left| \hat{D}(R_{i_c j_c}^{k_1}) - \hat{D}(R_{i_c j_c}^{k_2}) \right| \tag{12}$$

where $\hat{D}(R_{i_c j_c}^{k})$ is the average depth over the pixels $R_{i_c j_c}^{k} \subset \Omega_{i_c j_c}^{k}$. The sets of pixels $R_{i_c j_c}^{k-1}$ and $R_{i_c j_c}^{k-1}$ are obtained by finding the common contour $G_{i_c j_c}^{k_1 k_2}$ of $\Omega_{i_c j_c}^{k_1}$ and $\Omega_{i_c j_c}^{k_2}$, dilating $G_{i_c j_c}^{k_1 k_2}$ with radius of 5 pixels to get $\bar{G}_{i_c j_c}^{k_1 k_2}$ and taking the intersection so that $R_{i_c j_c}^{k_1} = \bar{G}_{i_c j_c}^{k_1 k_2} \cap \Omega_{i_c j_c}^{k_1}$ and $R_{i_c j_c}^{k_2} = \bar{G}_{i_c j_c}^{k_1 k_2} \cap \Omega_{i_c j_c}^{k_2}$. In this way we are able to merge regions with smooth transitions in depth on the contours by taking into account only a subset of the pixels in the region when computing the minimum of criterion (12). We notice that this approach yields better results in terms of final code length than considering whole regions. Results obtained with this partition are denoted with suffix D-RM in the method name. The presented methods for obtaining the center view partition are illustrated in Fig. 5.

B. Optimizing Partitions According to MDL Criterion

One downside to C and D segmentation methods is that they depend on parameters that are image dependent and not known a priori. In this section we propose a method that starts from arbitrary oversegmentations (e.g. a C-segmentation, obtained by mean shift algorithm, with small bandwidth parameters) and merges the regions incrementally according to the MDL criterion itself.

The partition minimizes the description length of data plus model. Here we discuss the MDL criterion that is not evaluating only the predictor efficiency, as we did in the case of sparse predictor design, but it encompasses also the cost incurred by the partition. This section presents our strategy and practical considerations for optimizing the sum of code-lengths for each term involved in the description of data and partition. The resulting partition is denoted in Section V with suffix CL-RM in the method name.

If we allowed an arbitrary segmentation (to prediction regions) for all the N^2 views, the overall code length associated with the encoding-decoding paradigm presented in Section II should be minimized by solving

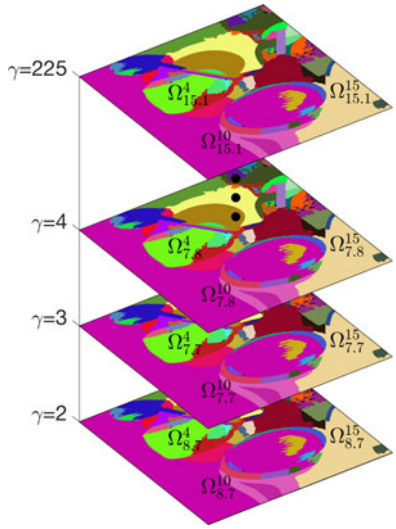


Fig. 6. Partition propagation. All $N^2 - 1$ partitions $\mathcal{P}_{i,j} = \{\Omega_{i,j}^k\}$ where $i = 1, \ldots, N$, $j = 1, \ldots, N$ (here we refer to the pair with a single index γ as in Fig. 1) emerge from the center view partition $\mathcal{P}_{i_c, j_c}$. The partitions are shown in pseudo color and a few example regions are labeled for each view.

$$\min_{\{\Omega_{ij}^k\}, \{\boldsymbol{\Theta}_{ijc}^k\}} \sum_{i,j,k,c} \left(\mathcal{L}_\varepsilon(\Omega_{ij}^k, \boldsymbol{\Theta}_{ijc}^k) + \mathcal{L}_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_{ijc}^k) \right) + \sum_{i,j} \mathcal{L}_{\mathcal{P}}(\mathcal{P}_{ij}) \tag{13}$$

where $\mathcal{L}_\varepsilon(\Omega_{ij}^k, \boldsymbol{\Theta}_{ijc}^k)$ is the cost of the c-component residuals over the region Ω_{ij}^k, when encoding with the optimal sparse predictor $\boldsymbol{\Theta}_{ijc}^k$. The term $\mathcal{L}_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_{ijc}^k)$ is the code length for encoding the optimal sparse predictor and $\mathcal{L}_{\mathcal{P}}(\mathcal{P}_{ij})$ is the cost of transmitting the partition $\mathcal{P}_{ij}$ for the view (i, j). The codelength over a region Ω_{ij}^k is denoted $\mathcal{L}(\Omega_{ij}^k) = \sum_c \left(\mathcal{L}_\varepsilon(\Omega_{ij}^k, \boldsymbol{\Theta}_{ijc}^k) + \mathcal{L}_{\boldsymbol{\Theta}}(\boldsymbol{\Theta}_{ijc}^k) \right)$.

Building completely distinct partitions over each view and sparse predictor (one for each color component and for each region of each partition) is in practice intractable, and because of that we resorted in the first place to propagate the partition of the central view to the side views using the horizontal and vertical displacements, so that the last term in (13) is

$$\sum_{i,j} \mathcal{L}_{\mathcal{P}}(\mathcal{P}_{ij}) = \mathcal{L}_{\mathcal{P}}(\mathcal{P}_{i_c j_c}) + \sum_{i,j,k} \mathcal{L}_\delta(\delta_{i,j}^k, \mu_{i,j}^k), \tag{14}$$

where the term $\sum_{i,j,k} \mathcal{L}_\delta(\delta_{i,j}^k, \mu_{i,j}^k)$ represents the cost of transmitting the disparities for the regions, and since we encode $\delta \in \Delta$ and $\mu \in \Delta$ with a fixed number of bits, see (1), the term is simply proportional with the number L of regions in the central partition. The result of partition propagation is visualized for the test image *Vespa* in Fig. 6. There, it can be seen that for example the center view region $\Omega_{8,8}^4$ has its counterpart in all the other $N^2 - 1$ views.

We proceed by tackling the optimization of (13) using a region merging strategy. The bit cost of (14) is very small compared to the bit cost of predictors so we need to locally optimize the term $\sum_{i,j,k} \mathcal{L}(\Omega_{ij}^k)$. The process is iterative and at each stage we choose two regions $\Omega_{ij}^{k_1}$ and $\Omega_{ij}^{k_2}$, we design MDL-optimal predictors for each of them as in Section III (all the

steps are the same e.g. the way of selecting regressors from available neighboring views) and we compute the MDL criterion using (8), cumulating over all color components. Then we repeat the process for the merged region $\Omega_{ij}^{k_1} \cup \Omega_{ij}^{k_2}$ and choose to merge the regions if the merging criterion evaluated over a set of views C':

$$\mathcal{J}_m(k_1, k_2) = \sum_{(i,j)\in C'} \mathcal{L}(\Omega_{ij}^{k_1} \cup \Omega_{ij}^{k_2}) - \left(\mathcal{L}(\Omega_{ij}^{k_1}) + \mathcal{L}(\Omega_{ij}^{k_2})\right) \tag{15}$$

is negative, $\mathcal{J}_m(k_1, k_2) < 0$. The set C' can contain all combinations of (i, j) or a subset of them, as discussed next. Our merging process works such that regions for which $\mathcal{J}_m$ is smallest are merged first. The computation of the criterion is then repeated for the resulting regions. The process is continued until all $\mathcal{J}_m > 0$ i.e. the merging of regions does not lead to improvement in final code length.

In evaluation of the criterion (15), we only consider such regions $\Omega_{ij}^{k_1}$ and $\Omega_{ij}^{k_2}$ that have a common contour. For example, in Fig. 6, the criterion (15) is evaluated for regions Ω_{ij}^{10} and Ω_{ij}^{4} but not for Ω_{ij}^{10} and Ω_{ij}^{15}. However, the process allows the regions to grow such that in the end even the regions Ω_{ij}^{10} and Ω_{ij}^{15} could belong to the same large region.

To optimize the overall code length, the set $\{\Omega_{ij}^{k}\}$ in (13) contains regions from all the $N^2 - 1$ views $(i, j) \in \{(i, j) \mid i = 1, \ldots, N,\ j = 1, \ldots, N,\ (i, j) \neq (i_c, j_c)\} = C$. We demonstrate in Section V (in particular in Figs. 10 and 11) that in order to ease computation time significantly without major impact on the final code length, we can choose only a subset of representative views so that $(i, j) \in C' \subset C$ where we can control the complexity with the cardinality of C'. This means that the first summation in (13) leads to similar partition result $\mathcal{P}_{i_c j_c}^{k}$ whether all the available views (i, j) are used or only a subset of them (see Fig. 10). For example, in Fig. 6, we could compute the merging criterion using only the regions in the view with $\gamma = 4$ which means that the merging process is based only on whether the criterion (15) leads to merging for regions $\Omega_{7,8}^{4}$ and $\Omega_{7,8}^{10}$ whereas the regions $\Omega_{7,7}^{4}$ and $\Omega_{7,7}^{10}$ play no role in the merging process.

The CL-RM method is summarized in Algorithm 6. When presenting results in Section V we append the cardinality of C' in the method's name i.e. a result using ten views is denoted CL-RM-10.

Algorithm 6: Code length region merging (CL-RM).

1: *Input:* 5D light-field L_F, subset of view indices C'.
2: Use mean shift algorithm to obtain initial over segmentation $\mathcal{P}_{i_c j_c}$ for the center view $A_{i_c j_c}$.
3: Find region displacements δ_{ij}^{k} and μ_{ij}^{k} for all regions Ω_{ij}^{k}, $(i, j) \in C'$, $k = 0, \ldots, L-1$ according to (1).
4: Obtain initial partitions $\mathcal{P}_{ij}$ for side views using Algorithm 1.
5: Find all neighboring regions $\Omega_{ij}^{k_1}$ and $\Omega_{ij}^{k_2}$, $(i, j) \in C'$ and compute the criterion $\mathcal{J}_m$ using (15).
6: **while** True **do**
7: Find $(k_1, k_2) = \arg\min_{k_1', k_2'} \mathcal{J}_m(k_1', k_2')$.
8: **if** $\mathcal{J}_m(k_1, k_2) > 0$ **then**
9: Go to step 15.
10: **else**
11: Merge regions k_1 and k_2 i.e. update accordingly both $\mathcal{P}_{ij}$, $(i, j) \in C'$ and $\mathcal{P}_{i_c j_c}$ and concatenate rows of matrices $\boldsymbol{D}$ corresponding to regions with indices k_1 and k_2.
12: Mark all $\mathcal{J}_m(k_1, k_2)$ to a big positive value and recompute (15) for all regions that were neighbors of either region with index k_1 or k_2.
13: **end if**
14: **end while**
15: *Output:* Merged center view partition $\mathcal{P}_{i_c j_c}$.

V. Experimental Results

In this section, we compare the proposed compressor's performance to existing lossless compressors and study the effect of different segmentation methods (discussed in Section IV) for center view partition and the choice of sparse prediction algorithm. We also demonstrate the usefulness of the MDL region merging algorithm discussed in Section IV-B.

For the experimental results, we use data from EPFL Light-Field Image Dataset , that contains raw data captured with a Lytro Illum camera. In particular, we select the subset of twelve images that was used in the recent challenge at ICME 2016 and has become the standard testbed for light-field compression. The images are listed in Table I. We use the L_F-structures that were obtained at EPFL using the LFToolbox [31] and stored as Matlab files for the ICME challenge[1] by the challenge organizers.

TABLE I
Test Images

i_I	Image Name
1	Bikes
2	Danger_de_Mort
3	Flowers
4	Stone_Pillars_Outside
5	Vespa
6	Ankylosaurus_&_Diplodocus_1
7	Desktop
8	Magnets_1
9	Fountain_&_Vincent_2
10	Friends_1
11	Color_Chart_1
12	ISO_Chart_12

A. Ranking Partitioning Methods and Sparse Prediction Algorithms

To demonstrate simple partitioning methods, we show for one of the data-set image, called *Bikes*, the result of using the tree different methods from Section IV-A for obtaining the partition with a similar number of regions in Fig. 5. The constant disparity segmentation (D) splits big uniformly colored regions into

[1]We buffered the used `.mat`-files at http://www.cs.tut.fi/∼helinp/slfc constructed as additional information page, supporting the reproducibility of our results.

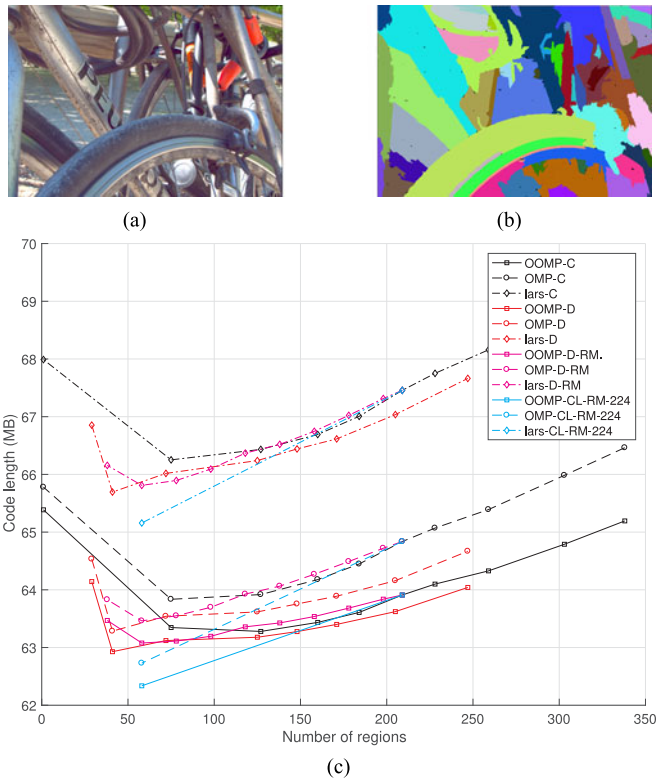


Fig. 7. Compression results in terms of total code length for test image *Bikes* with different partitioning methods and different sparse recovery algorithms. The best result is achieved by OOMP-CL-RM (lowest blue square of all plots) which used OOMP sparse design and region merging based on MDL. The segmentation obtained by CL-RM where the best code length is obtained is shown in (b). (a) *Bikes* image. (b) *Bikes* best segmentation. (c) Compression results for *Bikes*.

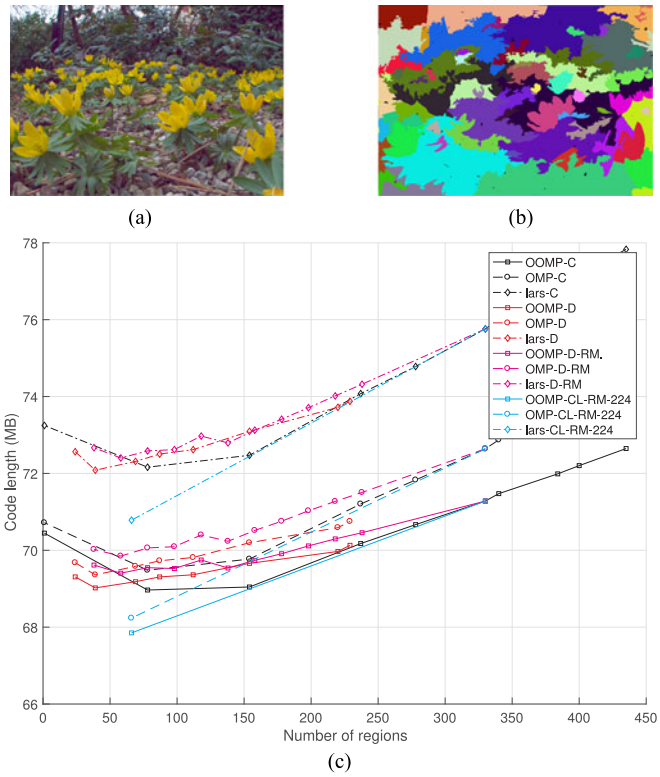


Fig. 8. Compression results in terms of total code length with different partitioning methods and different sparse recovery algorithms for *Flowers*. (a) *Flowers* image. (b) *Flowers* best segmentation. (c) Compression results for *Flowers*.

several separate regions if they are tilted towards the camera (see e.g. the tyre of the bicycle), while the color segmentation (C) has more precise object contours. The partition D-RM separates better the scene into objects, by using the depth information in addition to color homogeneity.

In Figs. 7 and 8 we show the compression performance with several parameters for C, D and D-RM and CL-RM segmentations as well as for all three sparse prediction methods. The points for C-segmentation are obtained using mean shift with bandwidth parameters $h_s = h_r = 4, 5, 6, 7, 8, 10, 12, 15$. The segmentation with only one region is manually set. The points for D-segmentation are obtained by uniformly quantizing the disparity map in $N_d = 3, 5, 8, 11, 14, 17, 20, 23$ levels and extracting the connected components. D-RM and CL-RM start from the C-segmentation with parameter $h_s = h_r = 8$ and proceed to reduce the number of regions by merging. The partitions were saved at certain points of this path and the full compression scheme was run for that particular partition to obtain the points in curves labeled D-RM and CL-RM.

We observe in Fig. 7, that in this case, as was also observed for all files, the best sparse recovery algorithm is OOMP, which is able to select a small number of regressors that enables the smallest prediction residuals with the smallest number of non-zeros in the sparse vector $\mathbf{\Theta}$. A similar result for the plenoptic image *Flowers* is given in Fig. 8.

Fig. 9 illustrates how the partitioning and sparse prediction algorithms perform. The comparison of partitioning methods for all images in terms of code length is given in Fig. 9(a). In most cases, the best code length is obtained with partitioning method CL-RM, as is expected since it is designed to optimize the code length under the assumptions presented in Section IV-B. For some test images, the constant depth regions provide an improvement in code length which implies that the sparse coding method is indeed more efficient when the regions are carved taking into account the depth of pixels. For most images, however, the color segmentation (C) provides the best results out of the methods without merging. For images with indices 6 and 8, the constant depth regions provided a better result than CL-RM because the C-segmentation from which CL-RM started had less regions than the winning D-segmentation.

The comparison of the sparse design algorithms is shown in Fig. 9(b) for the case of best C-segmentation. OOMP is always the best choice, followed by the algorithm OMP that was used in [12].

In region merging based on MDL (CL-RM), the effect of taking into account different numbers of neighbor views, as discussed in Section IV, is illustrated in Fig. 10 and the effect on processing time and performance is shown in Fig. 11. In Fig. 11 we selected the subset of view indices was selected equidistantly between views $\zeta = 10$ and $\zeta = 150$. It is also noticeable that the best result given in Fig. 9 a was found manually by running the compressor with different parameters of mean shift algorithm,

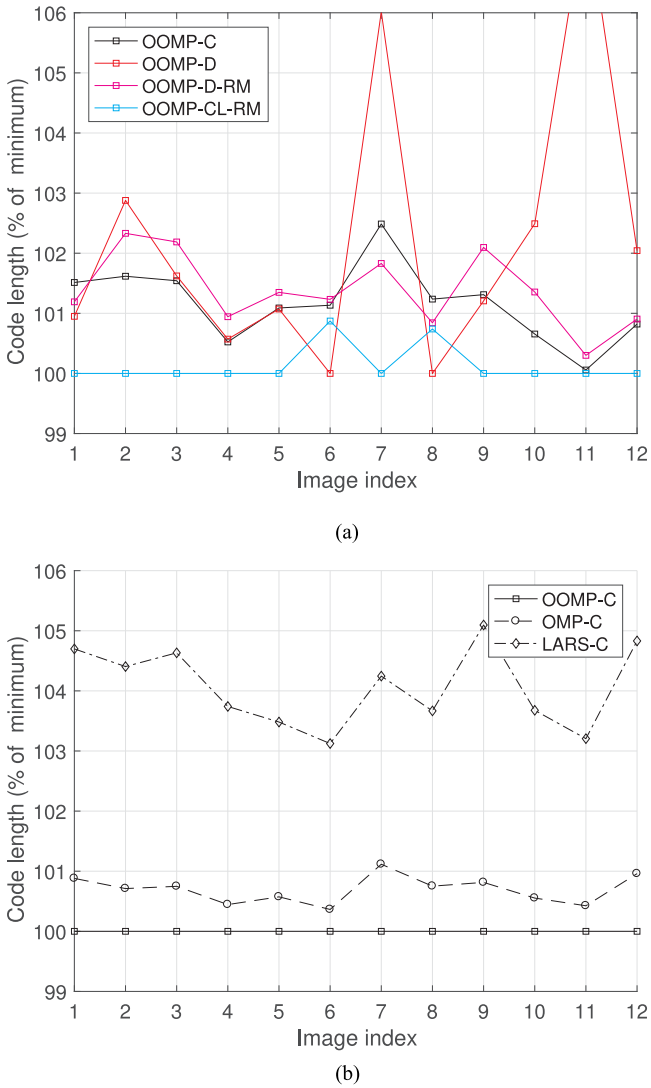


Fig. 9. Code length for each image in the dataset. (a) shows the best code length obtained with different partitioning methods while (b) compares the sparse recovery algorithms for the C-partitioning method. The results for different combinations of partitioning/sparse algorithms follow the trend shown by these results. (a) Partition methods. (b) Sparse recovery algorithms.

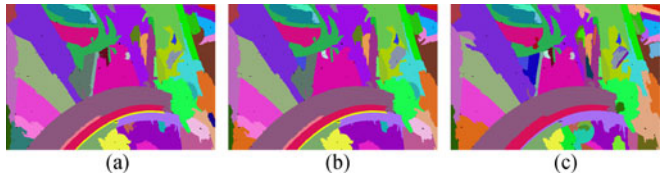


Fig. 10. The resulting segmentations from CL-RM process of MDL region merging, when using for computing the code length the predictors with various numbers of views for conditional coding. We notice that reducing the number of views results in a similar segmentation, which shows the stability of the process, irrespective on the number of views. (a) CL-RM result with 224 views. (b) CL-RM result with 18 views. (c) CL-RM result with one view.

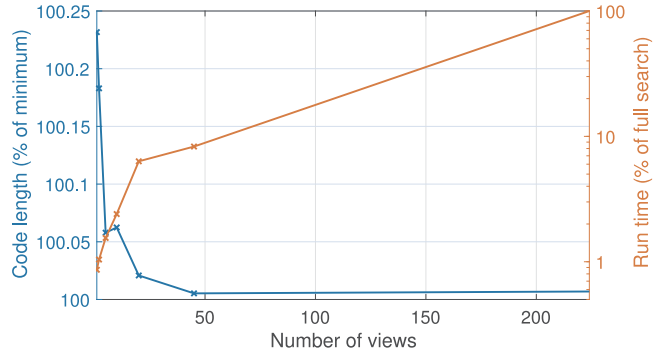


Fig. 11. The effect of utilizing different number of views, when computing the code length for CL-RM region merging. The horizontal axis represents the parameter $|C'|$. The average values of code length and processing time are shown over all 12 test images. The best code length is obtained with all 224 views. Increase in code length when less views are used in CL-RM is small but saving is significant in terms of processing time.

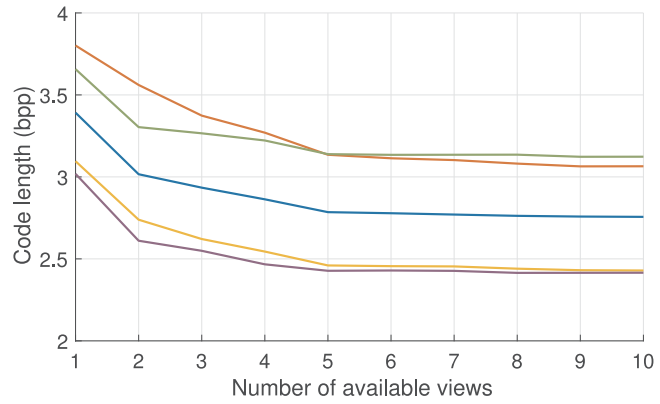


Fig. 12. Code length for the five largest regions in *Bikes* when the number of available views as regressors pixels is varied. After five views there is typically not much gain in code length but the complexity of encoding and decoding increases.

which was done solely to find the best performance of such a method. What we recommend instead is the usage of region merging, where no parameter tuning is needed. The best result after the region merging is obtained at once, which underlines its usefulness in practice. The effect of region merging is to combine regions that can share the same predictor, which typically leads to a much smaller code length for encoding predictors and slight improvement for encoding residuals since the order that the MDL-criterion selects is naturally higher when encoding a larger region.

We study the effect of adding more than 5 views to the set Γ_{ij} in Fig. 12. The number of columns in the regression matrix $\mathbf{D}$ increases by nine if we increase the number of used side views by one. This has two effects: a slower computation of the optimal sparse model, and also an increase in the codelength needed by the sparsity mask; on the other hand, when using an additional side view the quality of the prediction may increase, and hence the codelength of the residual may be smaller. The code length for a few example regions for various numbers of side views is shown in Fig. 12. Experimentally we limited the number of views to five.

We observe that OOMP leads to smallest predictor orders on average followed by OMP and then LARS. This is demonstrated in Fig. 13 where we show the average order of all predictor

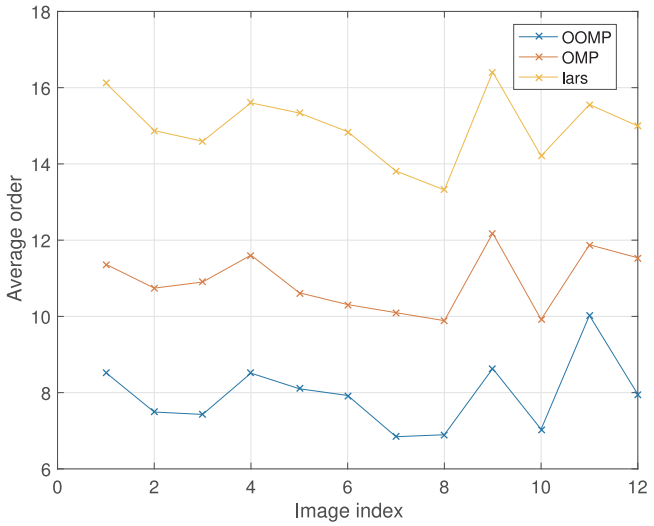


Fig. 13. Average order selected by the methods. Typically, the obtained code length for residuals, $\mathcal{L}_\varepsilon$ is similar for all methods, but the methods are getting those residuals with different predictor orders, leading to different code lengths for predictors $\mathcal{L}_{\boldsymbol{\Theta}}$, and finally different overall code length, $\mathcal{L}_\varepsilon + \mathcal{L}_{\boldsymbol{\Theta}}$. The distribution of orders for the image with index 1 can be seen in Fig. 14.

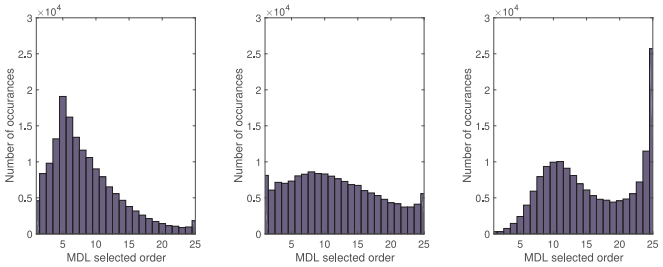


Fig. 14. The distribution of the order selected by the MDL criterion for different methods, for *Bikes* and a C-segmentation. There are in total 161211 predictor design problems involved in this case. The algorithms were run until $K = 25$ regressors were selected. (a) OOMP. (b) OMP. (c) LARS.

search problems over the test images. A closer view into the distribution of the orders is shown in Fig. 14.

B. Comparison to Existing Compressors

Here, the performance of the proposed compressor in terms of final code length is compared to existing generic tools for compressing L_F losslessly.

We use two methods for re-organizing the L_F data structure.

1) The views can be treated as a pseudo-video-sequence using the same spiral (see Fig. 1) sequence as for our method. Video compressors such as HEVC or AVC can be used.
2) The views can be rearranged in a single image I such that $I((x-1)N+i,(y-1)N+j,c) = A_{ij}(x,y,c)$, where $i = 1,\ldots,N$, $j = 1,\ldots,N$, $x = 1,\ldots,\xi$, $y = 1,\ldots,\eta$ and $c = 0,1,2$ [32]. This structure can be efficiently compressed with image compression tools such as JPEG 2000, CALIC [18] or FP8 compressor[2].

The final results are shown in Table II. There, we refer with HEVC and AVC to their ffmpeg (version N-82833-gedb4f5d) implementation. We aim to maximize their performance by using the preset option *placebo*. The method HM refers to HEVC Test Model, which we ran with the *random access*-preset. The exact configuration files are given in http://www.cs.tut.fi/∼helinp/slfc.

TABLE II
COMPRESSED FILE SIZES IN MEGA BYTES (MB)

i_I	HEVC	HM	AVC	CALIC	FP8	[12]	OOMP-CL-RM-1
1	81.82	88.32	80.61	82.47	82.99	69.66	**62.85**
2	87.04	95.46	86.36	88.63	86.75	77.06	**67.06**
3	86.65	96.25	86.19	88.65	86.06	78.14	**68.18**
4	81.50	87.75	80.18	81.30	78.76	67.28	**61.82**
5	84.33	92.38	82.71	87.94	83.23	71.90	**66.20**
6	70.86	77.79	69.74	75.82	68.40	58.99	**53.63**
7	56.47	72.77	55.87	63.45	52.86	50.79	**47.29**
8	70.74	77.21	69.07	75.95	67.17	59.56	**54.54**
9	85.51	92.67	84.15	85.75	85.67	70.20	**64.26**
10	69.46	77.76	67.58	75.20	66.92	60.19	**53.51**
11	82.83	96.40	81.95	93.36	78.92	71.29	**65.53**
12	78.43	84.96	78.35	76.85	75.04	71.40	**61.57**

TABLE III
THE SPLIT OF THE FINAL CODE LENGTH INTO DIFFERENT ELEMENTS, PRESENTED AS PERCENTAGE

i_I	$\mathcal{L}_\varepsilon$	$\mathcal{L}_\gamma$	$\mathcal{L}_\theta$	$\mathcal{L}_{\mathcal{P}_{i_c j_c}}$	$\mathcal{L}_{\delta,\mu}$	$\mathcal{L}_{A_{i_c j_c}}$
1	97.40	0.30	1.39	0.01	0.09	0.81
2	98.01	0.18	0.88	0.01	0.09	0.82
3	97.40	0.28	1.35	0.02	0.13	0.82
4	98.36	0.13	0.59	0.01	0.04	0.87
5	98.07	0.21	0.92	0.01	0.06	0.72
6	98.97	0.06	0.25	0.00	0.01	0.70
7	97.30	0.35	1.48	0.02	0.11	0.73
8	98.49	0.17	0.62	0.01	0.04	0.69
9	97.72	0.25	1.18	0.01	0.09	0.76
10	97.86	0.22	1.00	0.01	0.09	0.82
11	98.23	0.23	0.89	0.01	0.03	0.61
12	98.16	0.21	0.83	0.01	0.04	0.74

The reported file sizes are containing all the necessary information for decoding. Successful lossless decoding was checked. The size of the raw 24 bit data is 183 MB (we use SI-units where M denotes one million and the raw size in bytes is $3 \times N^2 \times \eta \times \xi$) and the compressed sizes are given in Table II. Results for JPEG2000 or JPEG-LS are omitted because CALIC outperforms them for all the images. It can be seen that our coding scheme achieves for every image the best results, improving significantly the compression rate over the generic compressors. To obtain the segmentation we use the parameterless CL-RM method with only a single view, with index $\zeta = 10$.

The division of code length into different components for the OOMP-CL-RM-1 method is given in Table III. The columns from left to right refer to: residuals, sparsity mask, sparse coefficients, center view partition, displacements and center view. It can be seen that the residuals are the major component.

VI. DISCUSSION

A. Computational Complexity of Encoding

In the full encoding scheme, the search for an optimal sparse predictor, correspoding to Algorithm 4, is the most

[2]Available at http://mattmahoney.net/dc/text.html#1532

computationally extensive task. We discuss the computational complexity with different choices of sparse prediction algorithm in Algorithm 4.

Without steps 17 and 18, the number of multiplications in Algorithm 4, when using OOMP (FOLS) (see [25])

$$\frac{1}{2}n(m+1)(m+2) - 3K + \sum_{k=1}^{K}(m-k+7)(m-k+1). \tag{16}$$

In our application, $m \leq 68$, whereas typically the number of pixels in a region n is in the range of thousands, K being fixed at 25. As mentioned in [25] FOLS is very efficient for this type of underdetermined system.

Our implementation of OMP exploits QR decomposition in solving step 15 of Algorithm 4. The number of multiplications when using OMP in Algorithm 4 is (see [24])

$$nm + \sum_{k=1}^{K}((n-k)(m+5) + (2m+1)k - m + 1) + 2(n-K+1) + (2m+1)(K-1) + m + \frac{1}{2}K(K-1). \tag{17}$$

With the (almost constant) parameters $m = 68$ and $K = 25$ in this application, in whole range of n encountered, OOMP (16) has about 25 per cent more multiplications that OMP (17).

For LARS, the complexity is dominated by the least-squares fitting and in general is given as $O(m^3 + nm^2)$ [26]. The implementation that we use [27] exploits a cross-product matrix $D^T D$ to reduce the number of operations.

The steps 17 and 18 in Algorithm 4 and steps 4 and 5 in Algorithm 5 are not included in the formulas above. They are the same for all algorithms. Step 17 of Algorithm 4 requires $2k$ bit shifts per iteration k and step 18 requires in total $\sum_{k=1}^{K} k^2 + k + 8(k+2)$ multiplications when exploiting the cross-product matrix $D^T D$ in the computation of the squared norm of residuals with a quantized predictor as we do in all implementations. The computational cost of these two steps is very low compared to other steps in Algorithms 4 and 5.

The encoder and decoder are implemented with C++. The sparse algorithms OOMP, OMP and LARS, that we refer to in this section, use Matlab. However, OOMP was also implemented in C++ and is included as a part of the encoder. In practice, our Matlab implementation of OMP is a little slower than OOMP. With a 4 GHz Intel Core i7 processor, running Algorithm 4 with CL-RM segmentation (and *Bikes* image) takes 1.5 seconds per view (the whole plenoptic image has $N^2 - 1$ such views) using OOMP, 2.1 using OMP and 2.6 using LARS.

B. Computational Complexity of Decoding

In decoding, the extensive task of running Algorithm 4 is avoided and simply at each pixel (x, y), the predicted value $\hat{y}$ is obtained using $k^\star$ integer multiplications. To obtain the exact reconstruction y, the correction, i.e. residual ε is obtained using arithmetic decoding with few integer divisions and multiplications per pixel.

Our non-optimized implementation of the decoder takes about one minute per plenoptic image.

C. Differences From the Baseline Method of [12]

The method used in this paper differs from the initial conference paper [12] in several respects, which leads to better compression results.

1) In [12] the partition into regions of the central view was obtained by the quantization of a disparity map (corresponding with our D-segmentation). Here, we consider color- and depth-based segmentation methods, combination of them and present a novel region merging process driven by an MDL criterion balancing optimally the codelength for coding the partitioning with the codelength for residuals and sparse predictors. The obtained segmentation can be valuable per se, and is encoded along the bitstream, so it is a side benefit for the decoder.
2) In [12] only OMP as a sparse design method was considered, here we experimented with alternatives and found OOMP to perform consistently better.
3) In [12] the MDL criterion for the sparse predictor design considered a simpler encoding strategy for the prediction mask and for the coefficients, while here the encoding uses variable length coding, accounted as such in MDL. This leads to a better penalisation of the model cost. Yet, the presented MDL criterion here is computationally simpler to evaluate.
4) In [12] cross-color prediction was not considered.

VII. Conclusion

A method for lossless compression of light-field data was described in detail and two major components affecting its performance, partitioning and sparse prediction, were investigated in detail. Sparse prediction provides state-of-the-art compression results and thus suits well to the data. Out of the investigated sparse predictor design methods OOMP is the best, being a fast, yet effective method for finding the sparse predictors. We have shown that the method for partitioning has a big impact on the final compression and both the color and the depth provide useful information. We proposed a region merging approach for obtaining the partition that achieves the best compression results. The interaction between the partitioning method, the prediction templates, the predictor design and the MDL criterion is very complex and can be further studied for obtaining improved results. The lossless compression of plenoptic images is important in applications where data integrity has to be preserved, e.g., in medical imaging or data storage, and is also useful in later stages of processing since it is guaranteed not to introduce any unpredictable artifacts for applications such as photo refocusing or the obtaining of an extended depth-of-field image.

References

[1] G. Lippmann, "Épreuves réversibles donnant la sensation du relief," *J. Phys.*, vol. 7, no. 4, pp. 821–825, 1908.

[2] C. Conti, L. Soares, and P. Nunes, "HEVC-based 3D holoscopic video coding using self-similarity compensated prediction," *Signal Process.: Image Commun.*, vol. 42, pp. 59–78, 2016.
[3] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Scalable coding of plenoptic images by using a sparse set and disparities," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 80–91, Jan. 2016.
[4] C. Conti, P. Nunes, and L. D. Soares, "HEVC-based light field image coding with bi-predicted self-similarity compensation," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*, Jul. 2016, pp. 1–4.
[5] R. Monteiro *et al.*, "Light field HEVC-based image coding using locally linear embedding and self-similarity compensated prediction," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*, Jul. 2016, pp. 1–4.
[6] Y. Li, R. Olsson, and M. Sjöström, "Compression of unfocused plenoptic images using a displacement intra prediction," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*, Jul. 2016, pp. 1–4.
[7] C. Perra and P. Assuncao, "High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*, Jul. 2016, pp. 1–4.
[8] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudosequence-based light field image compression," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*, Jul. 2016, pp. 1–4.
[9] I. Viola, M. Rerabek, T. Bruylants, P. Schelkens, F. Pereira, and T. Ebrahimi, "Objective and subjective evaluation of light field image compression algorithms," in *Proc. 32nd Picture Coding Symp.*, Nuremberg, Germany, 2016, pp. 1–5.
[10] M. Rizkallah, T. Maugey, C. Yaacoub, and C. Guillemot, "Impact of light field compression on refocused and extended focus images," in *Proc. 24th Eur. Signal Process. Conf.*, Aug. 2016, pp. 898–902.
[11] C. Perra, "Lossless plenoptic image compression using adaptive block differential prediction," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 1231–1234.
[12] P. Helin, P. Astola, B. Rao, and I. Tabus, "Sparse modelling and predictive coding of subaperture images for lossless plenoptic image compression," in *Proc. 3DTV-Conf.: True Vision—Capture, Transmiss. Display 3D Video*, Jul. 2016, 1–4.
[13] S. Huq, A. Koschan, and M. Abidi, "Occlusion filling in stereo: Theory and experiments," *Comput. Vision Image Understanding*, vol. 117, no. 6, pp. 688–704, Jun. 2013.
[14] I. Tabus, I. Schiopu, and J. Astola, "Context coding of depth map images under the piecewise-constant image model representation," *IEEE Trans. Image Process.*, vol. 22, no. 11, pp. 4195–4210, Nov. 2013.
[15] Y. Morvan, D. Farin, and P. H. N. de With, "Depth-image compression based on an R-D optimized quadtree decomposition for the transmission of multiview images," in *Proc. Int. Conf. Image Process.*, 2007, pp. 105–108.
[16] I. Tabus and P. Astola, "Sparse prediction for compression of stereo color images conditional on constant disparity patches," in *Proc. 3DTV-Conf.: True Vision—Capture, Transmiss. Display 3D Video*, Jul. 2014, pp. 1–4.
[17] T. M. Cover, "Enumerative source encoding," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 1, pp. 7377, Jan. 1973.
[18] X. Wu and N. Memon, "Context-based lossless interband compression-extending CALIC," *IEEE Trans. Image Process.*, vol. 9, no. 6, pp. 994–1001, Jun. 2000.
[19] M. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS," *IEEE Trans. Image Process.*, vol. 9, no. 8, pp. 1309–1324, Aug. 2000.
[20] F. Ghido and I. Tabus, "Sparse modeling for lossless audio compression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 14–28, Jan. 2013.
[21] L. R. LaMotte and R. R. Hocking, "Computational efficiency in the selection of regression variables," *Technometrics*, vol. 12, no. 1, pp. 83–93, 1970.
[22] S. Chen, S. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identication," *Int. J. Control*, vol. 50, pp. 1873–1896, 1989.
[23] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.
[24] S. F. Cotter, R. Adler, B. D. Rao, and K. Kreutz-Delgado, "Forward sequential algorithms for best basis selection," *IEE Proc.—Vision, Image Signal Process.*, vol. 146, no. 5, pp. 235–244, Oct. 1999.
[25] S. Chen and J. Wigger, "Fast orthogonal least squares algorithm for efficient subset model selection," *IEEE Trans. Signal Process.*, vol. 43, no. 7, pp. 1713–1715, Jul. 1995.
[26] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.
[27] K. Sjöstrand, "Matlab implementation of LASSO, LARS, the elastic net and SPCA," 2005. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?3897
[28] T. C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Proc. Int. Conf. Comput. Vision*, Dec. 2015, pp. 3487–3495.
[29] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. Int. Conf. Comput. Vision*, Washington, DC, USA, 2013, pp. 673–680.
[30] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.
[31] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2013, pp. 1027–1034.
[32] A. Vieira, H. Duarte, C. Perra, L. Tavora, and P. Assuncão, "Data formats for high efficiency coding of Lytro-illum light fields," in *Proc. Int. Conf. Image Process. Theory, Tools Appl.*, Nov. 2015, pp. 494–497.

Petri Helin received the M.Sc. degree in information technology in 2013 from Tampere University of Technology, Tampere, Finland, where he is currently working toward the Ph.D. degree in signal processing. His research interests include sparse modeling, compression, and 3-D image processing.

Pekka Astola received the M.Sc. degree in signal processing and communications engineering in 2013 from Tampere University of Technology, Tampere, Finland, where he is currently working toward the Ph.D. degree under the supervision of Prof. I. Tabus. His research interests include stereo imaging and its applications and image compression.

Bhaskar Rao (F'00) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, in 1979, and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, CA, USA, in 1981 and 1983, respectively. Since 1983, he has been with the University of California, San Diego, CA, where he is currently a Distinguished Professor with the Department of Electrical and Computer Engineering. He received the 2016 Signal Processing Society Technical Achievement Award for fundamental contributions to array processing and sparsity-based signal processing.

Ioan Tabus (SM'99) received the M.S. degree in electrical engineering from "Politehnica" University of Bucharest, Bucharest, Romania, in 1982, and the Ph.D.(Hons.) degree from Tampere University of Technology (TUT), Tampere, Finland, in 1995. In 1996, he was a Senior Researcher and, since January 2000, has been a Professor with the Department of Signal Processing, TUT. His research interests include 3-D image processing, audio, image and data compression, and genomic signal processing. He has coauthored two books and more than 250 publications in the fields of signal compression, image processing, bioinformatics, and system identification. He was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING between 2002 and 2005 and the Editor-in-Chief of the *EURASIP Journal on Bioinformatics and Systems Biology* between 2005 and 2013.

Analyzing Interfaces and Workflows for Light Field Editing

Marta Ortín, Adrian Jarabo, Belen Masia, and Diego Gutierrez

***Abstract*—With the increasing number of available consumer light field cameras, such as Lytro, Raytrix, or Pelican Imaging, this new form of photography is progressively becoming more common. However, there are still very few tools for light field editing, and the interfaces to create those edits remain largely unexplored. Given the extended dimensionality of light field data, it is not clear what the most intuitive interfaces and optimal workflows are, in contrast with well-studied two-dimensional (2-D) image manipulation software. In this work, we provide a detailed description of subjects' performance and preferences for a number of simple editing tasks, which form the basis for more complex operations. We perform a detailed state sequence analysis and hidden Markov chain analysis based on the sequence of tools and interaction paradigms users employ while editing light fields. These insights can aid researchers and designers in creating new light field editing tools and interfaces, thus helping to close the gap between 4-D and 2-D image editing.**

***Index Terms*—Editing interfaces, editing workflow, light fields, user study.**

I. Introduction

LIGHT fields are four-dimensional representations of a scene, where the two extra dimensions code angular information about the scene being captured. Leveraging this additional wealth of information allows to create effects such as small parallax shifts, synthetic refocusing after capture, or even scene reconstruction (see [1] for a recent survey on light field imaging and its applications). With the introduction of light field cameras in the consumer market (such as Lytro [2], Raytrix [3] or the Pelican camera for mobile devices [4]), they are becoming an increasingly popular alternative to traditional 2D images.

Editing traditional 2D photographs is a well-understood process with an established workflow. Moreover, existing image editing programs share the main ideas around which their interfaces are built, such as working with layers, creating masks, or basic point-and-click operations. However, finding a user-friendly interface to edit light fields remains an open problem: it is not clear what the best way to edit the four-dimensional information stored in a light field is, both algorithmically, and from a user's perspective. Our objective is to examine how subjects perform precise edit placement (see Fig. 1). This is a key issue in user-assisted processing tools for light fields [5],[7], since processing tools often require user input of some kind; past light field processing methods have employed different techniques to do this, but it is unclear what the advantages and drawbacks of the different techniques are. Jarabo *et al.* [8] recently proposed the first comprehensive study on the topic: They evaluated a set of basic tools on the two most common light field interface paradigms: parallax-based, and focus-based. In their work, they asked participants to perform several edits on synthetic and real light fields, such as changing the color of an object, painting on its surface, or altering exposure (data was made publicly available by the authors[1]). As a result of the analysis, many valuable insights were provided on aspects such as the suitability of the provided interfaces and the irrelevance of inaccuracies in the depth information used, which is key to assert the appropriateness of using depth maps reconstructed from captured data.

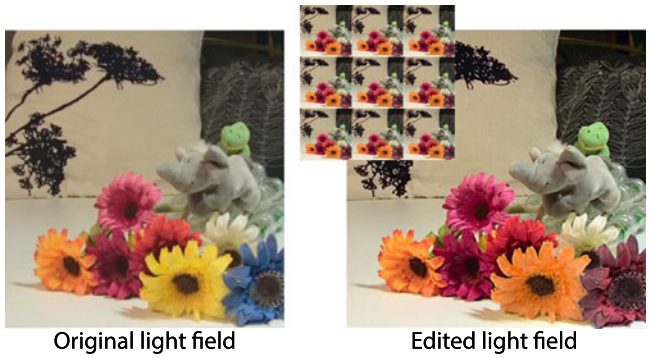


Fig. 1. Example result of a real light field captured with the Lytro camera and edited with our tool. In this work, we evaluate the benefits of different light field interaction paradigms and tools, and draw conclusions to help guide future interface designs for light field editing.

This paper presents a complementary analysis that focuses on another, but nevertheless important, aspect of editing: we analyze the subjects' preferred *workflows*. We examine the temporal domain of the editing process by performing detailed state sequence analysis and Markov chain analysis, including several metrics: mean time using each editing tool, tool usage distri-

Manuscript received December 29, 2016; revised June 13, 2017; accepted August 5, 2017. Date of publication August 29, 2017; date of current version October 23, 2017. This work was supported in part by the European Research Council under the European Union's Horizon 2020 research and innovation programme (CHAMELEON project) under Grant 682080 and in part by the Spanish Ministerio de Economía y Competitividad (TIN2016-78753-P and TIN2016-79710-P). The guest editor coordinating the review of this paper and approving it for publication was Prof. Jingyi Yu. *(Corresponding author: Marta Ortín-Obón.)*

The authors are with the Universidad de Zaragoza (I3A), Zaragoza 50018, Spain (e-mail: ortin.marta@unizar.es; adrianjarabo@gmail.com; bmasia@unizar.es; diegog@unizar.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2746263

[1]http://giga.cps.unizar.es/ ajarabo/pubs/lfeiSIG14/

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

bution, representative tool sequences, effective tool transitions, interface usage distribution, and effective interface transitions. For a number of typical scenarios (editing of surfaces, editing in free space, occlusion handling, and editing of complex geometries) we describe in detail the trends found, confirm the conclusions drawn in the previous work, and come up with novel insights.

This paper is an extended version of our previous publication on this topic [9]. Our main contribution is the methodology of the analysis we perform to assess subjects' editing workflows. Analyzing workflows is a challenging task due to the large inter- and intra-user variability, given the immense number of possible editing paths one may take. We model editing workflows as sequences of states, where states can either be the tools or the interfaces used, and apply state sequence analysis and Markov modeling. Our analysis is consistent with previous findings [8], [9], and offers additional insights, including:

1) depth information is largely used for editing when available; thus, interfaces with depth are preferred;
2) there is no clear or preferred order in the use of tools for any interface;
3) despite this, the workflow is a constant iteration of drawing/erasing and checking results, and only one or two representative sequences are required to model the behaviour of 35% to 90% of the subjects in most cases, indicating a high agreement among them;
4) the preferred form of visualization of the light field is that based on parallax, i.e., switching between neighboring views;
5) the preferred form for editing the light field is highly dependent on the task being performed.

This study provides a more comprehensive description of subjects' choices and preferences with regard to light field editing interfaces, together with a better understanding of the effort necessary to perform the different editing tasks. This combined knowledge is a valuable source of information for developing novel, user-friendly interfaces, as well as efficient and intuitive editing tools.

II. Related Work

A. Editing Tools

We focus here on light field editing, and refer the reader to the excellent survey by Schmidt *et al.* [10] for a more general discussion on visual content editing. Seitz and Kutulakos [11] performed edits such as painting and scissoring by creating a voxel-based representation of the light field to propagate the edits. Zhang *et al.* [12] and Wang *et al.* [13] performed morphing of two light fields, for which users need to indicate corresponding polygons between them to guide the process. User interaction was also required in the Pop-Up Light Field work by Shum *et al.* [14], in which users marked the silhouettes of objects in multiple views, which were then used to separate the light field into depth layers. Following the same idea, LightShop is a modular system to manipulate, composite, and render light fields [15]. Tompkin *et al.* [16] present an interaction system for 4D light field displays based on capturing the position and orientation of a pen. Nguyen *et al.* [17] edited the local light field at surfaces for material appearance editing in complex three-dimensional scenes. Jarabo *et al.* [5] devised a system for propagation of sparse edits between views, based on pixel affinity measured in the multidimensional light field space. A similar approach was later presented by Williem *et al.* [7]. Recently, a combined patch-based plus propagation approach was developed by Zhang and colleagues [18], capable of handling occlusions in tasks like object resizing or repositioning. Finally, Garces *et al.* [19] proposed a method to decompose a light field into its intrinsic components, allowing users to edit the albedo and lighting independently. All these works focus on the particular editing tools; in contrast, we concern ourselves with the interaction required by the user, involving both the interface used, and the workflows followed by the users.

B. Interfaces and Workflows

Many different interfaces and workflows have recently been explored for a wide range of applications, including authoring of 3D models [20], freeform design of objects [21], drawing lines to depict shapes [22], facial animation [23], sketching 3D surfaces [24], lighting design [25], [26], material editing [27], [28], icon selection [29], style similarity [30], [31], or simulation of cloth [32], to name a few. Closer to our approach, Santoni *et al.* [33] presented a statistical analysis of 3D digital sculpting. The authors analyzed users' behavior as they freely sculpted organic models to discover patterns in the data and trends in the workflows. Regarding the particular case of light fields, Jarabo *et al.* [8] presented the first thorough study to evaluate different light field editing interfaces from a user perspective. They identified the local point-and-click operation as the basis for many common edits (such as drawing, increasing brightness or placing additional objects), and performed two different experiments: In the first one, their goal was to compare two interfaces (based on parallax and focus cues, respectively) with respect to their effectiveness, efficiency and subjective preference, by performing several editing tasks of increasing complexity. The conclusions learned guided a second experiment, where users were given the possibility of combining both interface paradigms, and use new tools proposed by the users themselves.

In this paper, we build on this recent work and provide a statistical analysis of subjects' workflows and preferences for each type of editing scenario: editing of surfaces, editing in free space, occlusion handling, and editing of complex geometries. This complements previous work [8] by focusing on the user's behaviour on the temporal domain, that is, on editing *workflows*, which was previously ignored.

C. Signal Processing and Editing

Several works on image editing have used signal processing tools for applications such as tone mapping [34], [35], contrast enhancement [36], [37], relighting and recoloring [38], or material editing [39]. Many of these works can be coupled with manual intervention to guide the output of such techniques. For light field editing, a few works have directly used such processing tools: Most edit propagation techniques for light fields [5][7]

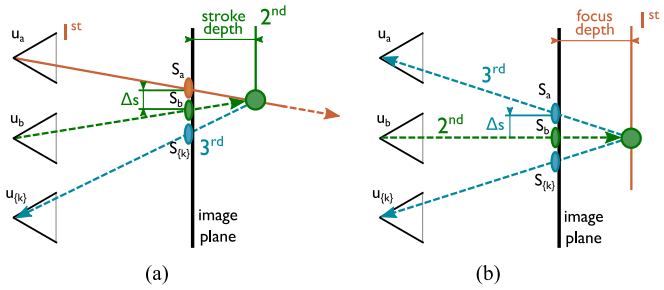


Fig. 2. Placing a stroke in the light field using our two interaction paradigms. (a) Multiview: The stroke s_a is first drawn in a view u_a (1st), and then its depth adjusted in another view u_b by displacing the stroke along the epipolar line (2nd), once this depth is chosen, the stroke is fixed and propagated to other views $u_{\{k\}}$. (b) Focus: First the depth is selected by adjusting the in-focus plane (1st), and a stroke s_b is placed in the central view u_b (2nd), which is then propagated to the other views $u_{\{k\}}$ (3rd).

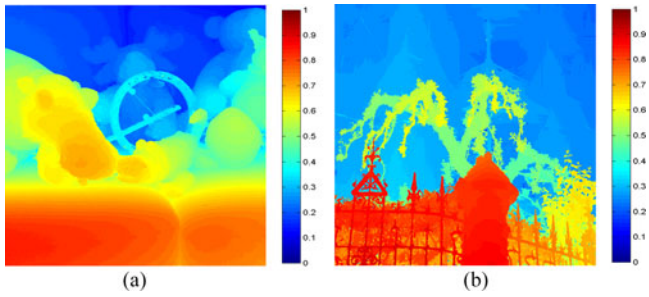


Fig. 3. Depth maps from light fields (central view) in the dataset by Kim *et al.* [46], used in the tasks of our second experiments (depth values are encoded as disparity in pixels). (a) Couch. (b) Mansion.

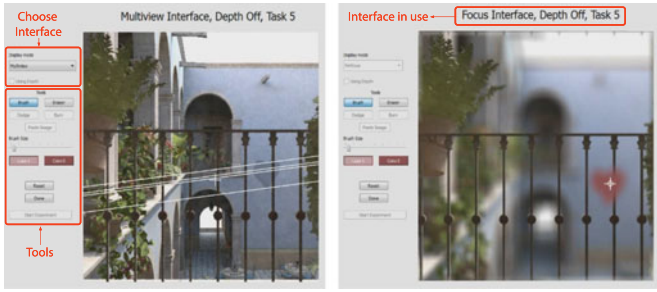


Fig. 4. User interfaces used in our tests. In the multiview paradigm, epipolar lines (white) mark the trajectory a currently active stroke will follow when moved in depth. In the focus paradigm, the in-focus plane marks the depth at which strokes will be placed. (a) Mutiview. (b) Focus.

are based on the bilateral filter when defining the propagation energy metric, while some commercial light field software incorporates light field-specific postprocessing filters (e.g., *Lytro Desktop* [2] or *Lightfield Iris* [40]). Finally, Garces *et al.* [19] expanded the classic Retinex constraint for intrinsic decomposition in light fields. However, in order to act locally or to allow user refinement, these works need precise edit placement; these local operations are the main focus of our work.

III. Editing Interfaces and Tools

A. Interaction Paradigms

The majority of previous works on light field editing rely on correspondences between the views to specify the desired edits (see for instance [11], [41], [42]). This amounts to using parallax as a depth cue to specify the depth at which an edit should be placed. We term this approach *Multiview (M)*. A completely different paradigm based on Isaksen *et al.'s* light field reparameterization [43] was explored by Davis *et al.* [44]; depth at which the editing will be performed is specified by a plane of focus, while the rest of the light field is blurred accordingly (wide aperture rendering). This amounts to using blur as a depth cue to specify an edit's position in depth, and we term this approach *Focus (F)*. Thus, both paradigms differ in how the *depth* at which the edit is to be placed is specified. These are the two paradigms originally explored by Jarabo and colleagues [8], which we also use for the analysis in this paper.

We now briefly summarize the procedure to perform an edit in the light field with each of our two basic interaction paradigms: Multiview and Focus. Note that we work with *point-based, local edits* (strokes). Our assumption is that a local point-based interaction lies at the basis of the majority of editing processes and tools (from the use of a brush or the eraser to more complex editing tools like selection or global filters). To place a stroke using the *Multiview* interface, the user first draws it in one view, and the epipolar lines of the stroke then become visible. By switching between views and observing the resulting parallax of the stroke, the user then infers its depth, and can move it to the desired position in the light field (Fig. 2, left). Using the *Focus* interface, the user first selects the plane of focus, then draws a stroke on that plane (Fig. 2, right). Notice the main difference in both interaction paradigms: in Multiview mode the stroke is first drawn, and then its position adjusted, while in Focus mode the depth is selected first, and then the stroke is drawn.

B. Interfaces

The first two interfaces we study are derived from the interaction paradigms described above, which we call *Multiview (M)* and *Focus (F)*. In addition to these, we leverage recent advances in scene reconstruction from light fields (e.g., [45], [46]), which allow to infer depth maps, and incorporate that information in two additional interfaces. Fig. 3 shows example depth maps for two of the light fields used in our tests. The availability of depth information for the input light field may alter the preferred workflow when users perform edits. As such, in the Multiview and Focus interfaces above, we include the option to use depth, yielding the *Multiview with Depth (MD)* and the *Focus with Depth (FD)* interfaces. For both, the strokes drawn will now snap to the nearest surface below them. Therefore, the main difference between these two interfaces lies in the visualization of the light field and the performed edits: While in Multiview with Depth these are visualized by switching the point of view, in Focus with Depth this is done by shifting the in-focus plane. All interfaces share the same screen layout, shown in Fig. 4. Details on the implementation of the interfaces can be found in Jarabo *et al.* [8].



Fig. 5. Target image and description for three representative editing tasks (from left to right, S1 and S3 from Experiment 1, and R5 from Experiment 2). For images of the remaining editing tasks, please refer to [8]. (a) Draw on wall. (b) Edit highlights. (c) Add ivy to the wall.

C. Editing Tools

Our interfaces incorporate the following set of basic tools: draw, erase, and pasting of pre-loaded billboard objects. To navigate the light field, they include: change view (only available in the Multiview interfaces) and set depth (only available in the interfaces without depth). In the Focus interfaces, depth is navigated using the mouse wheel. We further include a *depth selection* tool, which allows the user to pick a certain depth (mouse click), and specify a depth range around it using a slider. Edits can then only be performed inside the specified depth range; this is similar to a common mask tool, but working in depth space. Similar in spirit, a *color selection* tool is also included. With it, the user selects a color and a threshold, creating a corresponding selection mask. Finally, we incorporate a *visual aid* tool, which highlights the selected areas of the light fields by overlaying a semi-transparent checkerboard pattern on all other, non-selected areas.

For a better understanding of the interfaces and tools available we refer the reader to the videos of editing sessions which can be found online[2].

IV. Experimental Design

We briefly describe the two experiments carried out, whose results are thoroughly analyzed in the following sections. In every case, both objective and subjective data was collected. Objective data comprised timing data, tools used, or measures of error in depth when available. Subjective data included ratings and rankings by users collected via questionnaires and free form comments. More detailed information can be found in Jarabo *et al.* [8].

A. Synthetic Scenarios (Experiment 1)

1) Stimuli: We used two synthetic light fields for the different tasks: a complex architectural scene (*San Miguel*), and a still-life scene (*Vase*) (Fig. 5(a) and (b)). Since they are synthetic scenes, the actual ground truth geometry and depth are known.

2) Tasks: The first experiment contains five tasks [S1..S5], performed in fixed order. They include a target image, and subjects were given a detailed explanation of the desired edit. Tasks were chosen so that subjects would perform a variety of edits, including editing planar (S1) and curved surfaces (S2), emphasizing highlights (S3), placing objects in free space (S4), and dealing with occlusions (S5). Two examples are depicted in Fig. 5.

3) Procedure: Twenty paid subjects (6 female, 14 male) took part in the experiment. This is consistent with the number of subjects typically used in other similar statistical user studies, which usually range between two and 30 [24], [26], [27], [33]. Each subject used each of the four interfaces for all tasks, in random order to avoid possible learning effects. For each interface, tasks were performed in fixed order, S1 to S5. Time per task was limited to five minutes.

B. Real Scenarios (Experiment 2)

In Experiment 2, we used a hybrid interface combining all the different possibilities offered by the original four interfaces in Experiment 1, and used real, captured light fields. A key difference between the synthetic light fields used in Experiment 1 and real light fields is the absence of ground-truth depth information; instead, we are limited by the depth recovered by the camera, which is not error-free.

1) Stimuli: We used eight real light fields (an example is shown in Fig. 5(c)) captured by a Lytro light field camera, a camera gantry, or a Raytrix camera. They were chosen to cover a variety of scenes, depth, and reflectance complexities.

2) Tasks: This experiment contains ten tasks [R1..R10], in which a target image and a detailed explanation of the editing task to perform were given to subjects. The tasks were: colorization on slanted (R1) and curved surfaces (R2), cloning an object (R3), altering the appearance of a material (R4), adding texture (R5), importing a billboard object (R6), changing luminances (R7), altering color and placing a logo (R8), tweaking small details (R9), and harmonizing the color of the scene (R10). Again, tasks were chosen to cover a wide variety of edits. Fig. 5(c) shows the target image given to subjects for one of the tasks in Experiment 2.

3) Procedure: Ten different subjects (4 female, 6 male) took part in the experiment. Each subject was presented the ten tasks in random order to avoid possible learning effects. To carry out the tasks, they could choose freely between a Multiview and a Focus paradigm, and between using or not depth information. Time per task was limited to ten minutes.

V. Interface Suitability Based on the Task to Perform

This section discusses the suitability of each interface for the different editing tasks. We focus on Experiment 1, where we can easily compare how tasks are performed with each interface. Analysis of the data for time to completion, error in depth (measured as the L1 norm averaged across views), and ratings and rankings on interface preference provided by users yield three distinct clusters, roughly corresponding to three task categories: editing of surfaces (planar or curved), editing in free space, and occlusion handling. Data analysis was performed using repeated measures ANOVA for error, timings, and ratings, and Kruskal-Wallis for rankings [47].

[2]http://giga.cps.unizar.es/~ajarabo/pubs/lfeiSIG14/#downloads

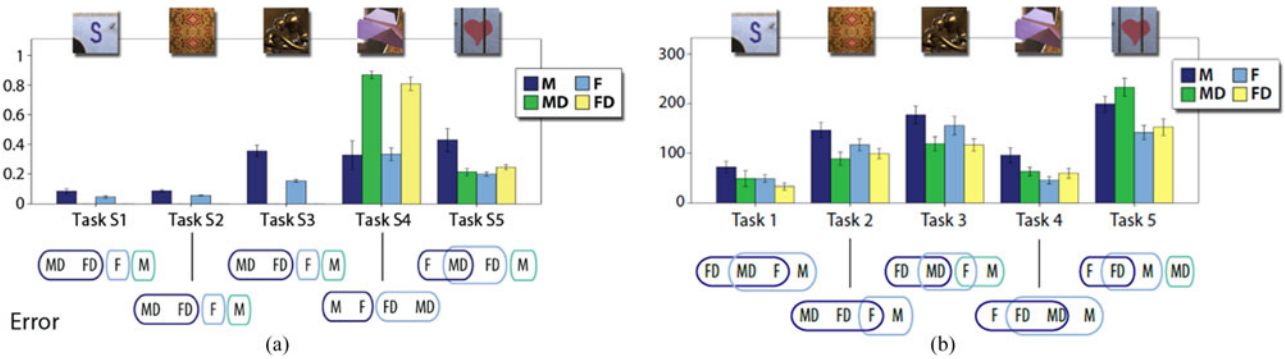


Fig. 6. **Left:** Mean error in depth for the five directed tasks in Experiment 1, and results of the pairwise comparisons between interfaces. **Right:** Same for time to completion. Source: [8].

A. *Editing of Surfaces*

Tasks S1 to S3 allow us to draw insights into surface editing. In these tasks, error in depth for interfaces with depth (MD and FD) is zero, since strokes snap to the surface below them (Fig. 6(a)). Consequently, realizing the task with these interfaces took less time (Fig. 6(b)). For interfaces without depth, M yielded a higher error than F, showing that users found it more difficult to locate an edit in depth with M. This is also reflected in the timings, in which M is significantly slower.

B. *Editing in Free Space*

Task S4 deals with positioning in free space. In this case, interfaces without depth yield lower error than those with depth, and the difference between interaction paradigms is not significant (Fig. 6(a)). Even though errors are high, times to completion are very low (Fig. 6(b)). In interfaces with depth, this can be due to the fact that users realized that they would not be able to correctly place it in depth (recall that the edit will snap to the surface right below) and gave up. This hypothesis seems supported by the low ratings these two interfaces received in this task. F is the interface of choice among the four tested for editing in free space, because it provides clear depth feedback and results in shorter editing times.

C. *Occlusion Handling*

Task S5 implies handling occlusions. For this, F yields both the lowest error and time to completion (Fig. 6), and its superiority is confirmed by ratings.

In summary, this experiment has shown the large influence of the task to perform in the interface of choice, with, essentially, MD and FD being ahead for on-surface editing, and F the interface of choice for free-space editing and occlusion handling.

VI. Workflow Analysis

Data collection from the experiments described in Section IV yields an immense quantity of both subjective and objective information, which provides insights into a variety of different aspects of light field editing: usage of different tools, preferred interfaces, variability of preferences with the task to perform, workflows, etc. While previous work has focused on how suitable different interfaces and tools are, and whether an interface can allow for satisfactory editing of light fields with current depth reconstruction methods [8], here we focus on subjects' workflows, that is, we look for underlying patterns in subjects' actions and their preferences for different generic tasks. As far as we know, this is an unexplored area of light field editing in which our intuitions on how users approach the editing tasks still need to be formally demonstrated.

Systematically analyzing the temporal activity of subjects while editing is a challenging task. We choose to model this data as categorical sequences, i.e., as a sequence of discrete possible states. States can either correspond to tools (Section VI-A) or to interfaces (Section VI-B), depending on what we want to analyze. There are multiple techniques in the literature to analyze categorical sequences, including correspondence analysis of states [48], or simply applying sequence alignment algorithms and computing pairwise distances [49]. In this work, we use several state sequence analysis techniques, including Markov modeling. We briefly explain here the fundamentals of these techniques, and then move on to describe the findings derived from them, both for tools and interfaces.

For the state sequence analysis—performed using the R library TraMineR [50]—, we first present visualizations of both individual and transversal statistics. The former refer to individual sequences, the latter present aggregated information, such as the state distribution plot, which shows, for each time instant, the aggregated distribution of states. More importantly, we then move on to compute *representative* sequences. One of the main concerns in state sequence analysis is to analyze the (dis)similarity between sequences and try to summarize subjects' behavior in a few possible sequences. A possible representative sequence is the medoid of the set, i.e., the sequence for which the sum of distances of the sequences in the set is minimized [50]. However, just one sequence does not adequately model the high variability present in cases like editing; we therefore compute a set of representatives, as it is explained below. Finally, to look into transitions we resort to Markov modeling [51]. We extract first-order Markov chains to analyze the transitions between states and determine which are the most common local workflows. Then, to further analyze more complex workflows,

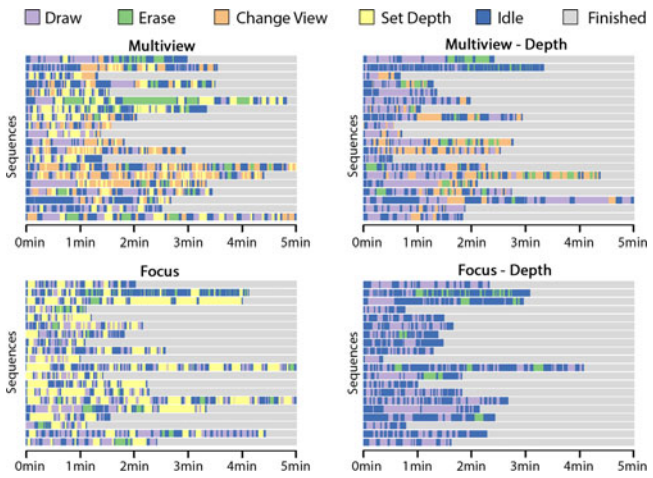


Fig. 7. Sequence index plot for Task S3 in Experiment 1. The horizontal bars show the sequence of tools the 20 participants used to complete the task with each one of the four interfaces, each tool being represented with a different color.

we search for higher-order hidden Markov sequences, which allows us to evaluate whether this local behavior holds over longer sequences.

In some cases, we only include graphs for some representative editing tasks. For the complete results, please refer to the supplementary document.

A. Tool Sequence Analysis (Experiment 1)

We assume that the tools used in the editing process correspond to different states: *draw*, *erase*, *change view*, and *set depth*. We include an *idle* state to represent that no tool is being used, and a *finished* state. We assume that idle periods shorter than one second correspond to switching from one tool to another; therefore, we eliminate them from the tool sequence. In this analysis, editing workflows are grouped by interface since it was preset and users could not switch among the four available ones.

1) Tool Sequence Visualization: Sequence index plots render state sequences and allow us to observe the succession of states and, through the length of each color segment, the time spent in each state. Fig. 7 represents the sequence of tools used by the 20 participants to complete Task S3 (edit specular highlights). While using M and MD, users constantly change view to check the appearance of the edits. This indicates that they are not approaching light field editing the way they approach conventional image editing; rather, they are very aware of the high dimensional nature of the scene. The use of change view is even more extensive when depth is not present, since they use it to place edits at a correct depth. This typically increases editing times: the time to complete the task tends to be longer when no depth is present. This holds for both F and M: a large amount of the time is spent on correct depth placement of edits (using the set depth tool to move the stroke along epipolar lines in M, or to set the desired plane in focus in F), which seems a strong advocate for the use of depth on light field interfaces. Interfaces featuring depth also register an increased time spent drawing (without an increase in erasing, which actually decreases, as we will show later in Fig. 8(a)), further indicative of the fact that users can concentrate more on the task at hand and its accurate completion.

2) Mean Time Using Each Tool: Fig. 8 shows the average amount of time, not necessarily consecutive, spent using each tool for Tasks S3 (edit specular highlights) and S4 (place object in free space). Note that, as explained before, the change view tool is only available in M, while the set depth tool can only be used when depth is off. Comparing Tasks S3 and S4 we also notice how different tools are involved in different tasks: while Task S3 requires drawing and erasing, in Task S4 the user only had to paste a new object, resulting in minimal drawing and erasing. At the same time, the comparison between such different tasks allows us to see patterns that are common despite the nature of the task. We observe that the time spent setting the depth is consistently longer in F. This is because modifying the depth is used first to choose the plane we want to edit, and then to check the result, which is costly for the user, resulting in longer editing times. In M, the depth is only used to adjust the stroke position. View changes are used to check results in both M and MD, and in the former also for stroke adjustment in depth. This generates a more frequent use of the change view tool for M compared to MD.

3) Tool Usage Distribution: A *state distribution plot* displays the general pattern of a whole set of state sequences. Unlike sequence index plots, they do not render individual sequences, but provide an aggregated view of the frequency of each tool for each time interval. Fig. 9 shows the tool usage distribution for all the tasks in Experiment 1. This represents how the use of different tools is distributed along time when varied editing tasks are performed. For example, in FD, at the beginning about 40% of the users are drawing while the rest are idle. After a while, a percentage of users erase strokes, and all the participants start to gradually finish the editing task.

Changing the view is more frequent when depth is not activated, since it is needed to adjust the position of strokes in different views, as opposed to MD, where it is only used to check results. Users spend more time in idle state in FD, probably because they are moving the cursor on the screen observing how the focus changes and deciding where to draw.

The main conclusion we extract from these plots is that there seems to be no clear or preferred order of states in any of the interfaces. The existence of a preferred pattern would mean, for instance, that at the beginning a majority draws, then a majority changes view to check results, then a majority erases mistakes, etc. However, the plots show that throughout time the distribution of the non-finished people among the available states remains constant, e.g, in MD, shortly after the beginning, there is a 1/3-1/3-1/3 distribution among draw-change view-idle that approximately remains true throughout time. However, we will see in the next paragraph that underlying patterns of actions can be found if shifting in time is considered when looking for them.

4) Representative Tool Sequences: Once we allow time shifting in the search for patterns, we find that a large number of editing sequences can be summarized with a representative set.

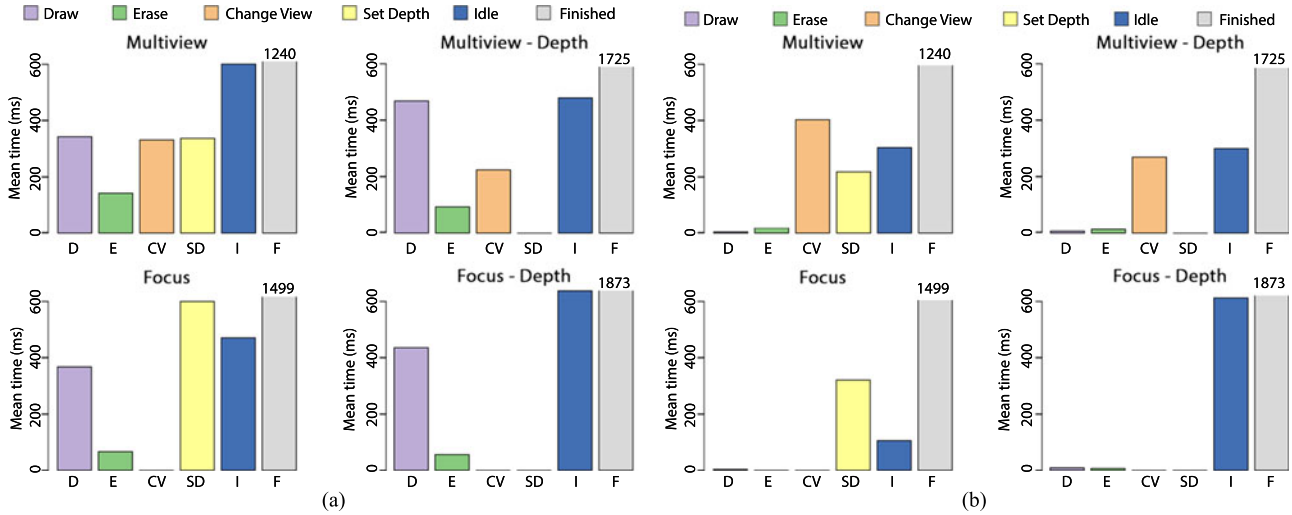


Fig. 8. Mean time using each tool for Tasks S3 and S4 in Experiment 1. Results are grouped by interface; in reading order: Multiview, multiview with depth, focus, and focus with depth. (a) Task S3. Edit specular highlights. (b) Task S4. Add airplane in free space.

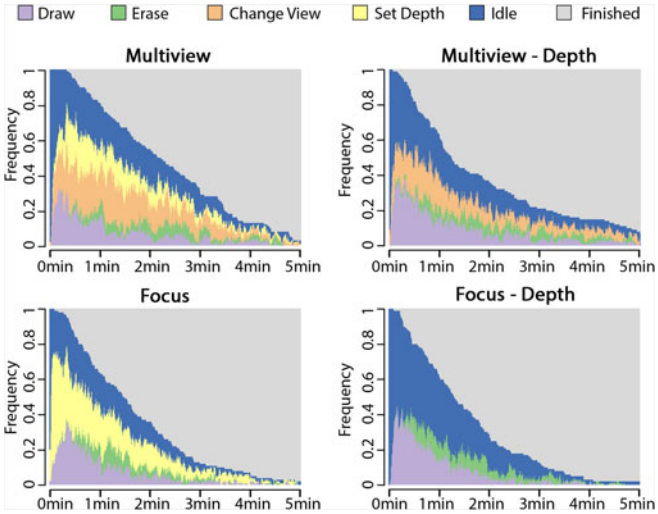


Fig. 9. Tool usage distribution for all tasks in Experiment 1 grouped by interface. Note that graphs do not represent individual sequences, they show the frequency of use of each tool at each time interval.

Fig. 10 shows the set of representative sequences for Tasks S1 and S3 grouped by interface. The sets cover at least 25% for the sequences in each one of the groups, meaning that the obtained sequences are representative of at least a quarter of the total number of sequences. In order to select the representatives, we compute the pairwise dissimilarities among sequences by calculating the Optimal Matching (OM) distance [52]. OM gives us the minimum cost of transforming one sequence into another allowing two transforming operations: the substitution of one element by another, and the insertion/deletion of an element, which generates a one-position shift of all the elements to its right. We have computed this metric with an insertion/deletion cost of one and a substitution cost matrix based on observed transition rates.

Comparing the representative tool sequences for a given task with the four interfaces, we can see clear differences. The sequence followed in FD is typically composed of idle and drawing intervals, while in MD, changing view is also used. It is interesting to note that, if depth is off, variability among users is much higher (Fig. 10, right); this is especially the case in M. This may indicate increased hesitation and experimenting: the path to completion is less clear, and users need to experiment more and end up following different paths.

Comparing two different tasks we notice that some of them can be edited by different users following many different sequences, while other (simpler) tasks are edited very similarly by all users: in Task S4, two to five representative sequences are required to obtain 25% coverage (Fig. 10, right); instead, in Task S1 only one representative covers 50% or more of the sequences (Fig. 10, left).

5) Effective Tool Transitions: Transition rates between each couple of states give us the probability of switching from one state to another. In image editing, users always spend long periods of time on the same task (e.g., drawing), and normally are idle for a while when they change from one tool to another. To avoid that behavior from biasing our results, we consider only *effective transitions*, that is, we eliminate self-loop transitions and remove the idle state. Fig. 11 presents the probability of the effective transitions for all the tasks in Experiment 1, grouped by interface.

In M (Fig. 11(a)), the editing workflow consists of constantly changing the view to observe the appearance of strokes and setting the depth to move them along the epipolar lines, with occasional drawing and erasing to perform small corrections. We also notice that users never finish editing directly after drawing, because they have to check how the new strokes look in different views first. In MD (Fig. 11(b)), strokes automatically snap to the surface below the cursor, so adjusting depth is not necessary. Participants change the view only to check their edits, and draw and erase to fix the errors they detect. Again, most transitions to finish come from change view, indicating that users use it to check the result.

With F (Fig. 11(c)), setting the depth is extensively used in combination with drawing and erasing. The large number of

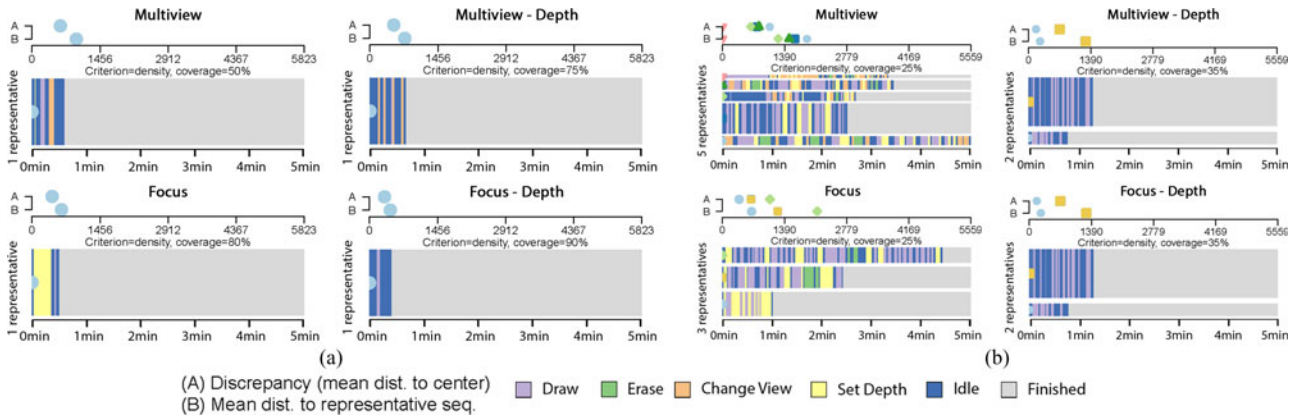


Fig. 10. Representative tool sequences for Tasks S1 and S3 in Experiment 1 grouped by interface with a minimum coverage of 25% and optimal matching distance used to compute dissimilarities. Representative sequences are plotted bottom-up according to their representativeness score and the bar width is proportional to the number of sequences assigned to them. Above each plot, two values are displayed for each representative: the discrepancy within the subset of sequences assigned to that representative (a), and the mean distance from the sequences to the representative (b). The scale ranges from zero to the maximum theoretical distance.

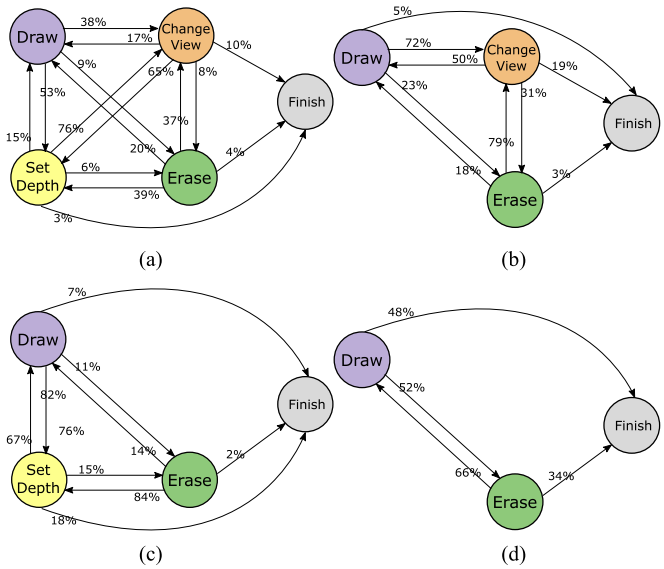


Fig. 11. Effective tool transition probabilities for all the tasks in Experiment 1. Self-loop transitions have been removed and the idle state is not considered. Note that set depth only appears when depth is not activated. With depth on, edits are automatically placed in the surface below the cursor. (a) Multiview. (b) Multiview with depth. (c) Focus. (d) Focus with depth.

transitions from draw and erase to set depth and from set depth to finish shows that users not only use it to set the desired plane in focus but also to check the results. In this case, the view cannot be changed. Finally, when using FD (Fig. 11(d)) neither change view nor set depth are available. The user must simply draw and erase until (s)he obtains the desired effect, and can check the results by moving the cursor on the screen to set the region below the mouse in focus.

In order to further analyze these workflows, we extract the hidden Markov chains (MC) up to order five (chains with six states), which complement the first order MC shown in Fig. 11. This analysis shows that, for M, the most common workflow in all tasks consists of looping between setting the edits' depth and navigating through the light field views to check whether the edit is correctly placed. Besides, users generally navigate before finishing the task. On the other hand, the limited navigation capabilities of F simplifies the workflows: in general, users loop between setting the depth and drawing. Similar to M, the users usually navigate through the light field by using the only available degree of freedom, which in this case is depth focus. In the cases where depth is available (MD and FD), the workflows are significantly simplified, and users focus mostly on surface editing (drawing, and then erasing for refining). However, in MD, users still make heavy use of navigation to check the positioning of the edits and their correct propagation across the light field views. Again, in most occasions users check the correctness of the edits by navigating before finishing the task.

Overall, we make two main observations: First, the workflow is a constant iteration of acting upon the scene (drawing/erasing) and checking the results, which is common among artists working with other media too [33]. Our interfaces thus seem to offer a valid approach for the usual artist workflow to be applied to light fields. Second, our interfaces allow users to leverage the high dimensionality of light fields during editing, instead of finding it cumbersome to navigate.

B. *Interface Sequence Analysis (Experiment 2)*

In this experiment, users can freely choose which interaction paradigm (Focus or Multiview) to use during the editing process, as well as whether they want to use depth information or not. Further, they can change among these interfaces as often as they want during the editing process. Additionally, in this experiment there are two new tools available to users, namely the color and the depth selection tools (please see Section III for detailed explanations). Therefore, we now include the use of these interfaces considering five states: *Multiview*, *Multiview with Depth*, *Focus*, *Focus with Depth*, and *finished*. As mentioned above, our focus here is on workflows, so we look at times of use of the different interfaces and tools as an indicator of preferences. We also look into transitions between states, and additionally collect data on subjective preferences.

We cluster the tasks in categories (editing of—planar or curved—surfaces, editing in free space, and occlusion handling), with an additional category for editing objects of intricate

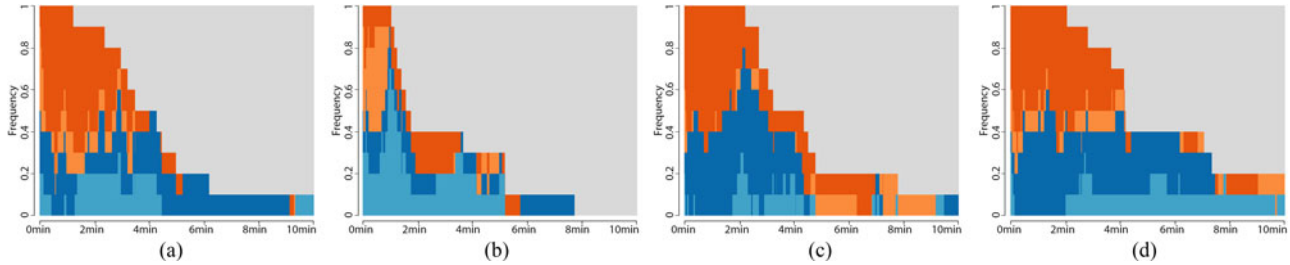


Fig. 12. Interface usage distribution for Tasks R1, R3, R5, and R10, one pertaining to each of the four categories or use case scenarios (editing surfaces, editing in free space, handling occlusions, and editing complex geometries). Note that graphs do not represent individual sequences, they show the frequency of use of each interface at each time interval. (a) Task R1 *(editing surfaces)*. (b) Task R3 *(editing in free space)*. (c) Task R5 *(handling occlusions)*. (d) Task R10 *(complex geometries)*.

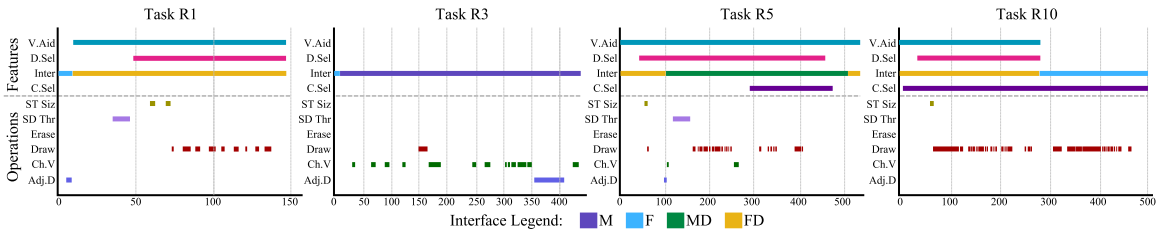


Fig. 13. Sample workflows for Tasks R1, R3, R5 and R10, one pertaining to each of the four editing categories (editing surfaces, editing in free space, handling occlusions, and editing complex geometries). They indicate the tools and interfaces used by the subject along time, shown in the abscissa (in seconds). The six bottom marks in the y-axis correspond to different operations; from bottom to top: set depth (*Adj.D*), change view (*Ch.V*), draw, erase, set depth threshold (*SD Thr*), and set tool size (*ST Siz*). The top four marks indicate whether that feature was activated or not at each time instant, and correspond, from bottom to top, to color selection tool (*C.Sel*), interface (*Inter*), depth selection tool (*D.Sel*) and visual aid tool (*V.Aid*). Note that *Inter* features different colors specifying the interface being used. These workflows for all subjects and tasks can be found at http://giga.cps.unizar.es/˜ajarabo/pubs/lfeiSIG14/.

geometry, for which subjects follow a different workflow than when editing simpler planar or curved surfaces. We choose four tasks as an example of each one of the four categories: R1, R3, R5, and R10, respectively.

1) Interface Usage Distribution and Sample Workflows: Tasks R1, R2, R7, R9, and R9 involve editing planar and curved surfaces. In all of these tasks, the use of depth information is largely favored, which matches our results in Experiment 1 and can be seen in the representative example depicted in Fig. 12(a). There is no clear preference between interaction paradigms (F or M), although there is a slight trend towards F; in the debriefing interviews, subjects reported that Focus offered a very strong and easy-to-interpret cue for visualization of the active area. Regarding the tools used, Tasks R2 and R9 favor the use of the color selection tool, possibly because the areas requiring editing are small, similar in color, and without a distinct depth with respect to their surrounding areas. The rest favor the use of the depth selection tool. A sample editing workflow for Task R1 is shown in Fig. 13.

Task R3 requires editing in free space. In this case, again, results are consistent with Experiment 1: depth information is scarcely used (Fig. 12(b)). We observe here a trend towards Multiview, as shown also in the workflow for Task R3 included in Fig. 13. This is possibly due to the absence of high frequency information in the area to edit, which causes the blur of the Focus interface to provide little or no depth cues.

In this experiment, Tasks R5 and R6 require dealing with occlusions. Here, the introduction of the depth and color selection tools causes a change with respect to results obtained in Experiment 1. While in the first experiment there was a large amount of erasing to deal with the occlusions, the introduction of the depth selection tool, largely used in both R5 and R6, reduces the need to erase to a minimum (see Fig. 13, Task R5, for a sample editing workflow in that task). Surprisingly, there is little difference between the use of MD and FD, revealing that as long as depth information and related tools are present, the interaction paradigm is less relevant for these tasks. The color selection tool is fairly used in Task R5 to avoid the pipe, which is hard to disambiguate from the rest of the wall in the depth dimension.

Tasks R4 and R10, which involve editing complex geometries, clearly show the need for the color and depth selection tools. When intricate geometries are present, these are extensively used. The nature of the scene determines which one is used: in the case of R4, 9 out of 10 subjects used the depth selection tool to complete the task, while in the case of R10, 9 out of 10 used the color selection tool, as shown in the sample editing workflow for Task R10 shown in Fig. 13. The majority of the subjects used depth information throughout the tasks; however, differences between the time of use of FD and MD were not significant (Fig. 12(d)).

2) Effective Interface Transitions: We analyze the probability of transitioning from one interface to another without considering self-loop transitions. Note that there are different user controls for setting Multiview or Focus, and depth or no depth. Therefore, it is not possible to transition directly from M to FD and vice versa, or from F to MD and vice versa. Fig. 14 shows the interface transition probabilities for the four representative tasks from Experiment 2: R1, R3, R5, and R10.

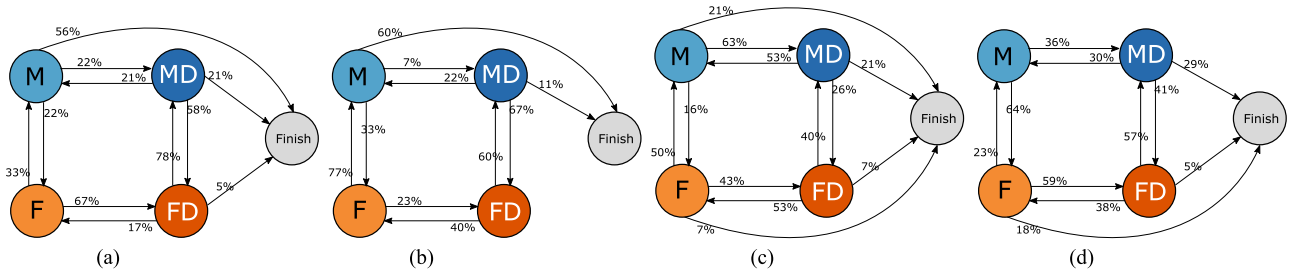


Fig. 14. Effective interface transition probabilities for Tasks R1, R3, R5, and R10, one pertaining to each of the four categories or use case scenarios (editing surfaces, editing in free space, handling occlusions, and editing complex geometries). Self-loop transitions have been removed. (a) Task R1 (*editing surfaces*). (b) Task R3 (*editing in free space*). (c) Task R5 (*handling occlusions*). (d) Task R10 (*complex geometries*).

In Task R1 (Fig. 14(a)), toggling between MD and FD is very common. This points out that users prefer the use of depth, and change between Multiview and Focus to find the most suitable interface to perform their edits or to check their results. In Task R3 (Fig. 14(b)) users always finish their editing in Multiview (with or without depth). This indicates that they prefer to use Multiview to verify that they placed the object correctly in free space, which concurs with our previous observations. For Tasks R5 (Fig. 14(c)) and R10 (Fig. 14(d)), which involve dealing with occlusions and complex geometries, the new depth and color selection tools determine the workflow, decreasing the importance of the choice of interface. This results in participants switching among interfaces looking for what feels more comfortable for drawing and checking results without any clear patterns on the choice of interface.

In summary, the second experiment confirms the findings of Experiment 1 in most aspects, with a clear exception in occlusion handling and intricate geometries, which are now easily dealt with thanks to the new tools. We also observe that depth information is almost always required, while the differences between the interfaces (Multiview and Focus) become less significant. Still, the Multiview paradigm (with or without depth) is the prevailing choice to examine the results, as shown by the fact that the majority of transitions to the Finish state tend to come from this paradigm.

VII. Conclusion

We have tested a set of interfaces and tools for light field editing, and shown that they allow users to satisfactorily edit light fields and perform common, everyday tasks, such as those in Experiment 2, on real, captured, light fields. While previous work has focused on the feasibility of using the light field editing interfaces, tools, and depth reconstruction methods, we present here the main findings in terms of workflow and preferences for the different scenarios that a user may encounter (free space, planar surfaces, occlusions, etc.). In order to do so, we have performed state sequence analysis and hidden Markov chain analysis focusing on both the editing tools and interaction paradigms.

With this new analysis, we have noticed that users quickly understand the high dimensionality of light field images and work on a constant iteration of drawing/erasing and checking the results by navigating across light field views or adjusting the depth. We have also demonstrated that having depth information available allows faster and more accurate editing and is preferred by users. When using that depth information, placing edits at the desired depth is straightforward, which reduces the need to use tools that select views or depth, and narrows their use to result checking. Another key finding is that users generally switch to the Multiview paradigm (with or without depth) to review their work before finishing.

Although users' actions and preferences in this study may be partially affected by the particular design of the light field editing tools, which may be rapidly updated by the community, we believe this work contributes to the foundations and provides a solid basis for the development of future tools and interfaces, both for researchers and UI designers. Further, we believe these insights can also be used for processing and editing other high dimensional data, such as BRDFs (bidirectional reflectance distribution functions) or BTFs (bidirectional texture functions), for which data-driven editing techniques remain an open problem [28], [53].

Acknowledgment

The authors would like to thank the anonymous reviewers for their insightful comments, and the members of the Graphics and Imaging Lab for their help.

References

[1] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.
[2] Lytro, "The lytro camera," 2014. [Online]. Available: www.lytro.com
[3] Raytrix, "The raytrix camera," 2014. [Online]. Available: www.raytrix.de
[4] K. Venkataraman *et al.*, "PiCam: An ultra-thin high performance monolithic camera array," *ACM Trans. Graph.*, vol. 32, no. 6, 2013, Art. no. 166.
[5] A. Jarabo, B. Masia, and D. Gutierrez, "Efficient propagation of light field edits," in *Proc. 5th Ibero-Amer. Symp. Comput. Graph.*, 2011, pp. 75–80.
[6] H. Ao *et al.*, "Light field editing based on reparameterization," in *Pac. Rim Conf. Multimedia*, 2015, pp. 601–610.
[7] Williem, K. W. Shon, and I. K. Park, "Spatio-angular consistent editing framework for 4d light field images," *Multimedia Tools Appl.*, vol. 75, no. 23, pp. 16615–16631, 2016.
[8] A. Jarabo, B. Masia, A. Bousseau, F. Pellacini, and D. Gutierrez, "How do people edit light fields?" *ACM Trans. Graph.*, vol. 33, no. 4, pp. 146:1–146:10, 2014.
[9] B. Masia, A. Jarabo, and D. Gutierrez, "Favored workflows in light field editing," in *Proc. Comput. Graph., Vis., Comput. Vis. Image Process.*, 2014.

[10] T.-W. Schmidt, F. Pellacini, D. Nowrouzezahrai, W. Jarosz, and C. Dachsbacher, "State of the art in artistic editing of appearance, lighting, and material," in *Proc. Eurographics, State Art Rep.*, 2014, pp. 216–233.
[11] S. M. Seitz and K. N. Kutulakos, "Plenoptic image editing," *Int. J. Comput. Vis.*, vol. 48, no. 2, pp. 115–129, 2002.
[12] Z. Zhang, L. Wang, B. Guo, and H.-Y. Shum, "Feature-based light field morphing," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 457–464, 2002.
[13] L. Wang, S. Lin, S. Lee, B. Guo, and H.-Y. Shum, "Light field morphing using 2d features," *IEEE Trans. Vis. Comput. Graph.*, vol. 11, no. 1, pp. 25–34, 2005.
[14] H.-Y. Shum, J. Sun, S. Yamazaki, Y. Li, and C.-K. Tang, "Pop-up light field: An interactive image-based modeling and rendering system," *ACM Trans. Graph.*, vol. 23, no. 2, pp. 143–162, 2004.
[15] D. R. Horn and B. Chen, "Lightshop: Interactive light field manipulation and rendering," in *Proc. Symp. Interact. 3D Graph. Games*, 2007, pp. 121–128. [Online]. Available: http://doi.acm.org/10.1145/1230100.1230121
[16] J. Tompkin *et al.*, "Joint 5d pen input for light field displays," in *Proc. 28th Annu. ACM Symp. User Interface Softw. Technol.*, 2015, pp. 637–647. [Online]. Available: http://doi.acm.org/10.1145/2807442.2807477
[17] C. H. Nguyen, D. Scherzer, T. Ritschel, and H.-P. Seidel, "Material editing in complex scenes by surface light field manipulation and reflectance optimization," *Comput. Graph. Forum*, vol. 32, no. 2pt2, pp. 185–194, 2013. [Online]. Available: http://dx.doi.org/10.1111/cgf.12038
[18] F.-L. Zhang, J. Wang, E. Shechtman, Z.-Y. Zhou, J.-X. Shi, and S.-M. Hu, "Plenopatch: Patch-based plenoptic image manipulation," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 5, pp. 1561–1573, May 2017.
[19] E. Garces, J. I. Echevarria, W. Zhang, H. Wu, K. Zhou, and D. Gutierrez, "Intrinsic light fields," *Comput. Graphics Forum*, 2017.
[20] H.-T. Chen *et al.*, "History assisted view authoring for 3D models," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2014, pp. 2027–2036.
[21] T. Igarashi, S. Matsuoka, and H. Tanaka, "Teddy: A sketching interface for 3D freeform design," in *Proc. SIGGRAPH*, 1999, pp. 409–416.
[22] F. Cole *et al.*, "Where do people draw lines?" *ACM Trans. Graph.*, vol. 27, no. 3, pp. 88:1–88:11, 2008.
[23] J. C. Miranda, X. Alvarez, J. Orvalho, D. Gutierrez, A. A. de Sousa, and V. Orvalho, "Sketch express: A sketching interface for facial animation," *Comput. Graph.*, vol. 36, no. 6, pp. 585–595, 2012.
[24] R. Schmidt, A. Khan, G. Kurtenbach, and K. Singh, "On expert performance in 3D curve-drawing tasks," in *Proc. Eurograph. Symp. Sketch-Based Interfaces Model.*, 2009, pp. 133–140.
[25] W. B. Kerr and F. Pellacini, "Toward evaluating lighting design interface paradigms for novice users," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 26:1–26:9, 2009.
[26] O. Karlík, M. Růžička, V. Gassenbauer, F. Pellacini, and J. Křivánek, "Toward evaluating the usefulness of global illumination for novices in lighting design tasks," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 6, pp. 944–954, Jun. 2014.
[27] W. B. Kerr and F. Pellacini, "Toward evaluating material design interface paradigms for novice users," *ACM Trans. Graph.*, vol. 29, no. 4, 2010, Art. no. 35.
[28] A. Serrano, D. Gutierrez, K. Myszkowski, H.-P. Seidel, and B. Masia, "An intuitive control space for material appearance," *ACM Trans. Graph.*, vol. 35, no. 6, 2016, Art. no. 186.
[29] L. F. Laursen *et al.*, "Icon set selection via human computation," in *Pacific Graphics Short Papers*, E. Grinspun, B. Bickel, and Y. Dobashi, Eds., The Eurographics Association: Goslar, Germany, 2016.
[30] E. Garces, A. Agarwala, D. Gutierrez, and A. Hertzmann, "A similarity measure for illustration style," *ACM Trans. Graph.*, vol. 33, no. 4, 2014, Art. no. 93.
[31] E. Garces, A. Agarwala, A. Hertzmann, and D. Gutierrez, "Style-based exploration of illustration datasets," in *Multimedia Tools Appl.*, vol. 76, no. 11, pp. 13067–13086, Jun. 2017.
[32] L. Sigal *et al.*, "A perceptual control space for garment simulation," *ACM Trans. Graph.*, vol. 34, no. 4, 2015, Art. no. 117.
[33] C. Santoni, C. Calabrese, F. Di Renzo, and F. Pellacini, "Sculptstat: Statistical analysis of digital sculpting workflows," arXiv:1601.07765, 2016.
[34] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, 2002.
[35] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 257–266, 2002.
[36] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski, "Edge-preserving decompositions for multi-scale tone and detail manipulation," *ACM Trans. Graph.*, vol. 27, no. 3, 2008, Art. no. 67.
[37] S. Paris, S. W. Hasinoff, and J. Kautz, "Local Laplacian filters: Edge-aware image processing with a Laplacian pyramid." *ACM Trans. Graph.*, vol. 30, no. 4, 2011, Art. no. 68.
[38] E. S. Gastal and M. M. Oliveira, "Domain transform for edge-aware image and video processing," *ACM Trans. Graph.*, vol. 30, no. 4, 2011, Art. no. 69.
[39] I. Boyadzhiev, K. Bala, S. Paris, and E. Adelson, "Band-sifting decomposition for image-based material editing," *ACM Trans. Graph.*, vol. 34, no. 5, 2015, Art. no. 163.
[40] Vertical Horizon, "Lightfield iris," 2013. [Online]. Available: http://www.verticalhorizon-software.com/LFI/index.html
[41] E. Stavrakis and M. Gelautz, "Image-based stereoscopic painterly rendering," in *Proc. Eurograph. Symp. Rendering*, 2004, pp. 53–60.
[42] W.-Y. Lo, J. van Baar, C. Knaus, M. Zwicker, and M. Gross, "Stereoscopic 3d copy & paste," *ACM Trans. Graph.*, vol. 29, no. 6, pp. 147:1–147:10, 2010.
[43] A. Isaksen, L. McMillan, and S. J. Gortler, "Dynamically reparameterized light fields," in *Proc. SIGGRAPH*, 2000, pp. 297–306.
[44] A. Davis, M. Levoy, and F. Durand, "Unstructured light fields," *Comput. Graph. Forum*, vol. 31, no. 2pt1, pp. 305–314, 2012.
[45] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4d light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 41–48.
[46] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 73:1–73:12, 2013.
[47] D. W. Cunningham and C. Wallraven, *Experimental Design: From User Studies to Psychophysics*. Boca Raton, FL, USA: CRC Press, 2011.
[48] J. C. Deville and G. Saporta, "Correspondence analysis with an extension towards nominal time series," *J. Econometrics*, vol. 22, pp. 169–189, 1983.
[49] A. Abbott and A. Tsay, "Sequence analysis and optimal matching methods in sociology, review and prospect," *Sociol. Methods Res.*, vol. 29, pp. 3–33, 2000.
[50] A. Gabadinho, G. Ritschard, N. Mller, and M. Studer, "Analyzing and visualizing state sequences in R with TraMineR," *J. Stat. Softw.*, vol. 40, no. 4, pp. 1–37, 2011.
[51] A. Berchtold and A. E. Raftery, "The mixture transition distribution model for high-order Markov chains and non-Gaussian time series," *Stat. Sci.*, vol. 17, pp. 328–356, 2002.
[52] A. Abbott and J. Forrest, "Optimal matching methods for historical sequences," *J. Interdiscip. History*, vol. 16, no. 3, pp. 471–494, 1986.
[53] A. Jarabo, H. Wu, J. Dorsey, H. Rushmeier, and D. Gutierrez, "Effects of approximate filtering on the appearance of bidirectional texture functions," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 6, pp. 880–892, Jun. 2014.

Authors' photographs and biographies not available at the time of publication.

Image Reshaping for Efficient Compression of Plenoptic Content

Xin Jin, *Senior Member, IEEE*, Haixu Han, *Student Member, IEEE*, and Qionghai Dai, *Senior Member, IEEE*

***Abstract*—With insertion of an additional microlens array in front of the imaging sensor, plenoptic camera, which acquires light field by a single exposure, attracts great interests both from academy and industry. A plenoptic image records both spatial and angular light radiance in a super high definition with macropixel structures, which results in a distinct intensity distribution. Several compression techniques have been proposed, however, most of them are suffered from the standard incompatibility and high compression complexity. In this paper, a novel and standard compatible plenoptic data reshaping algorithm is proposed to align the macropixel structures in the lenslet image with the coding unit grids of block-based image/video coding standards before compression. After encoding and decoding the reshaped lenslet image using the existing block-based image/video coding standards, the inverse image reshaping is used to convert the reconstructed image back to the original light-field data format for further applications. The experimental results reveal that the proposed method introduces 10.47% and 27.64% bitrate reduction to HEVC/H.265 format range extension profile and AVC/H.264 high 4:4:4 profile, respectively, by compressing the reshaped lenslet images. Also, it brings additional compression efficiency improvements to intraprediction tools, such as intrablock copy and locally linear embedding based coding method, by an average of 11.62% and 9.81% bitrate reduction as they compress the reshaped plenoptic images.**

***Index Terms*—AVC/H.264, HEVC/H.265, light field coding, plenoptic image compression, video coding.**

I. Introduction

RELYING on insertion of an additional microlens array in front of the imaging sensor, the so-called plenoptic camera, like *Lytro* [1], which acquires light field using a single exposure, attracts great interests both from academy and industry. Contrary to conventional cameras, the hand-held plenoptic camera can capture both spatial and angular light radiance in a single shot. The captured plenoptic images after post-processing can be applied to a variety of applications for improved quality of service [2]. For instance, plenoptic images can be projected onto an integral-imaging monitor for displaying 3D reconstructed images with full parallax [3], [4]; exploiting the focusness and depth information extracted, they can be utilized by saliency detection [5] for improve accuracy. Moreover, the subaperture images rendered from them can be exploited by object recognition and matching for improved efficiency [6]. Refocused panoramas can also be generated by the subaperture images employing pose estimation [7]. Therefore, plenoptic cameras are becoming increasingly popular in academy and industry.

An image captured by a plenoptic camera, which is referred to as a lenslet image in the following sections, is generally in super-high-resolution with a full color space for each pixel to preserve the fidelity of spatial and angular information. Additionally, each image consists of a number of macropixels, each of which gathers a group of pixels under a microlens. The pixels inside each macropixel records the light from different directions, which results in a distinct distribution relative to the pixels captured by conventional cameras. Also, the macropixel structure in the lenslet image changes the spatial correlations in the neighborhood. Consequently, efficient compression methods are highly desired for storing and transmitting the lenslet images efficiently.

The existing compression methods for lenslet images can be mainly classified into two categories: one is to compress the lenslet image directly and the other is to compress the light field data rendered from the lenslet image instead. The approaches that compress the lenslet image directly tried to improve the compression efficiency by exploiting the distinct distribution among the pixels. Some approaches used existing coding tools in JPEG, JPEG2000 and SPITH directly [8]. Novel coding tools like displacement Intra prediction [9], [10], Locally Linear Embedding (LLE) based intra coding mode [11] and disparity-compensated prediction [12], [13] were employed to exploit spatial correlation among the microlens images for improved compression efficiency. However, those methods introduce blocking artifacts to the decoded data due to the incoherence between macropixel structures in the lenslet image and the coding unit grids used by coding standards. Other techniques, such as 3D discrete wavelet transform (DWT) [14], compressive sensing [15], and statistical representation based on principal components analysis [16] were introduced to improve the compression efficiency. However, they cannot fully exploit the geometry information for each

Manuscript received December 25, 2016; revised June 4, 2017 and July 18, 2017; accepted August 5, 2017. Date of publication August 17, 2017; date of current version October 23, 2017. This work was supported in part by the NSFC Project, under Grant 61371138, China and in part by the Shenzhen Project, under Grant JCYJ20170307153135771, China. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Jingyi Yu. *(Corresponding author: Xin Jin.)*

X. Jin and H. Han are with Shenzhen Key Laboratory of Broadband Network and Multimedia, Graduate School at Shenzhen,Tsinghua University, Shenzhen 518055, China (e-mail: jin.xin@sz.tsinghua.edu.cn; haixu.han@qq.com).

Q. Dai is with the Tsinghua National Laboratory of Information Science and Technology and Department of Automation, Tsinghua University, Shenzhen 518055, China (e-mail: qhdai@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2741108

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

micro image and generally introduce high complexity in decoding.

Instead of compressing the lenslet image directly, approaches in the second category tried to compress the light field data rendered from the lenslet image instead. They generated a pseudo-temporal video sequence by extracting microlens images or subaperture images [17], [18], which generally correspond to the images from different viewpoints, from the lenslet image, and compressed them using the existing video coding standards, such as H.264 [19] and High Efficiency Video Coding (HEVC) standard [20]. Various scanning topologies [21], like raster scanning, spiral scanning, and parallel scanning, that reorder the subaperture images in the pseudo video sequence were proposed. Furthermore, Shasha *et al.* arranged subaperture images as the format of multi-view video and compressed it using multi-view video coding (MVC) standard [22]. Conti *et al.* proposed self-similarity compensated prediction [23] and geometry-based disparity compensation [24] to exploit the correlation between the subaperture images in the generated pseudo video sequence. To further exploit the disparity among the subaperture images, a sparse coding method [25] and a prediction method based on the statistical properties and geometry information [26] were proposed. Also, some transform-based methods [27]–[29] using discrete wavelet transform (DWT) were proposed to explore the correlation of different views. Although a much higher compression performance can be achieved compared with those approaches compressing lenslet images directly, they all suffer from hundreds of times increase in the computational complexity.

It is observed that due to the shape and arrangement of macropixels in the lenslet images, macropixels cannot be effectively partitioned into coding units (blocks) in a way that would be compressed efficiently by the existing block based compression standards. In our previous work [30], a preliminary plenoptic image compression scheme based on linear transformation and interpolation was proposed to align the macropixel structures with the coding unit grids of block based image/video coding standards. In this paper, we generalize the processes to be a lenslet image preprocessing scheme, called reshaping in the paper, which is parameterized universally to optical configurations of plenoptic cameras and the coding unit settings in the video coding standards, exploits the characteristics in the lenslet images and enables efficient compression of plenoptic content by using existing block based image/video coding standards. Comprehensive analyses are provided to demonstrate the efficiency in improving the spatial correlation among the neighboring prediction units, the attractive performance in improving the compression efficiency for HEVC/H.265 and AVC/H.264 together with a good tradeoff in computational complexity. Compared with the existing approaches, the major contributions of this paper are:

1) One of the first image reshaping approaches for lenslet images, which is friendly to the block based coding standards. The macropixel structures in the lenslet image are aligned with the coding unit grids, and the inherent correlation among macropixels can be efficiently exploited by the intra prediction methods;
2) The proposed approach is simple and compatible with any block based coding standard, like HEVC and H.264, with a significant improvement in the compression efficiency;
3) The proposed approach can work together with new intra prediction tools to bring additional compression efficiency improvements to lenslet image compression.

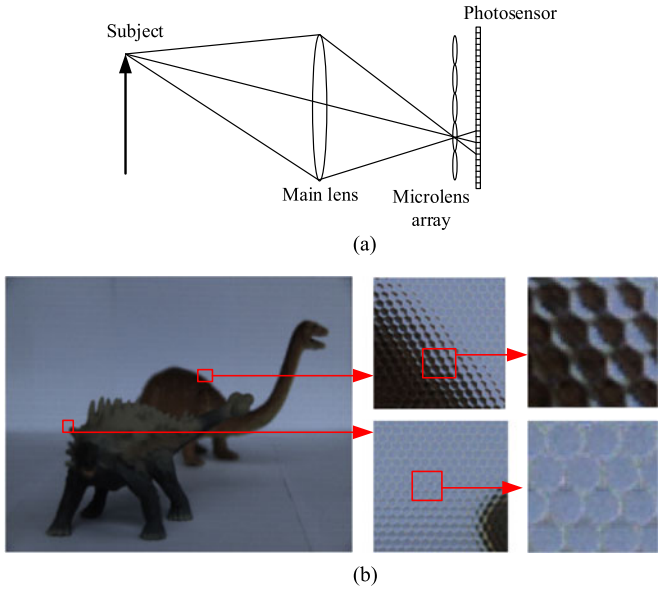


Fig. 1. (a) The architecture of a standard plenoptic camera; (b) A lenslet image [34] captured by *Lytro Illum* and its partial magnification.

The rest of the paper is organized as follows. The characteristics and preprocessing method of the lenslet image are introduced in Section II. The framework and the details of the proposed algorithm are described in Section III. Experimental results and analyses are provided in Section IV followed by conclusions drawn in Section V.

II. Lenslet Image Structure and Preprocessing

This section briefly describes the standard plenoptic camera system in terms of lenslet image structure and preprocessing for the captured raw lenslet images.

The architecture of a standard plenoptic camera is shown in Fig. 1(a), which inserts a microlens array at the image plane of main lens in front of the photosensor. The light rays coming from a focused spatial point go through the main lens and converge at a microlens. Then, they will be diverged by the microlens and be imaged by a group of pixels beneath the microlens to form a macropixel [31]. Thus, unlike a pixel in the image captured by conventional cameras recording the integrated intensity for the lights reaching the same position, a macropixel in a lenslet image consists of a group of angular pixels recording the radiance of the lights from different directions. An instance of a raw lenslet image captured by *Lytro Illum* [32] is illustrated in Fig. 1(b). As demonstrated by the magnifications, the lenslet image is composed of a number of macropixels hexagonal in shape, which record spatial and angular information. The number of macropixels is consistent with the number of microlens

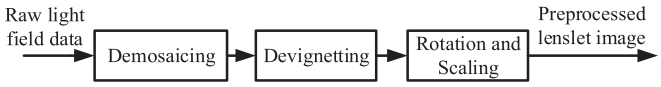


Fig. 2. The framework of preprocessing.

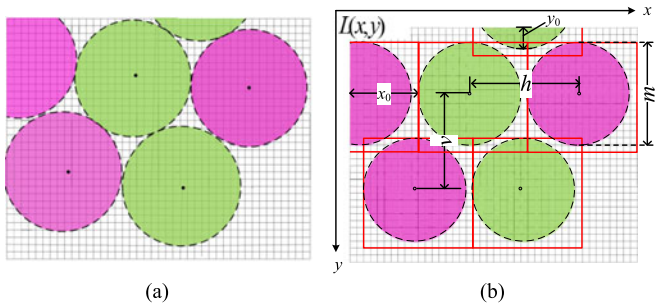


Fig. 3. (a) Vignetting-corrected lenslet image; (b) The preprocessed lenslet image.

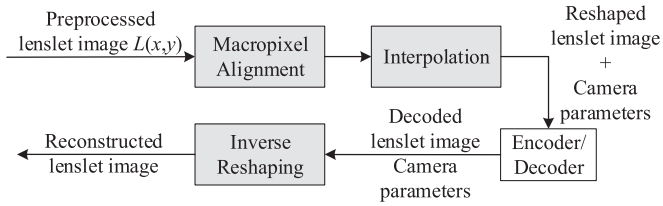


Fig. 4. Framework of the proposed method.

in the camera. Since the adjacent macropixels capture spatially connected area in a scene, a high geometry correlation and intensity similarity are preserved among the adjacent macropixels, especially when the object is homogeneous.

Since the exact placement of the lenslet array is always unknown and the optical parameters vary slightly because of the manufacturing diversity, the lenslets may be spaced by a non-integer multiple of pixel pitch with unknown translation and rotation. Hence, light field decoding of lenslet images is needed to align the hexagonally sampled data, corresponding to macropixels, to an orthogonal integer pixel grid, which is also recommended by JPEG PLENO [37] as a preprocessing technique for light field compression. Light field decoding, named by preprocessing in the following, can be divided into three steps: *Demosaicing*, *Devignetting*, *Rotation and Scaling* [33], as shown in Fig. 2:

Demosaicing is to reconstruct a color lenslet image with full RGB color space from the raw Bayer-pattern image with spatially undersampled color channels. The conventional linear demosaicing method is directly applied [33]. *Devignetting* is to correct vignetting effect. Vignetting results in the brightest spot in each white macropixel approximating the lenslet center and darker pixels appearing around the edge of macropixels. Devignetting corrects the effect by dividing the image by the white image.

Rotation and Scaling is to align macropixels to an orthogonal integer pixel grid. As shown in Fig. 3(a), since the lenslet spacing is a non-integer multiple of pixel pitch, the macropixels (i.e., circular areas in green or magenta) in the vignetting- corrected image are on a non-integer spaced, rotated grid relative to the image's pixels (i.e., grids in white). By selecting an appropriate white image provided by the camera, the grid parameters, such as translational parameters t_x and t_y, scaling parameters s_x and s_y, rotational parameters r_x and r_y, together with mean vertical and horizontal spacing and offset can be derived by locating the macropixel centers. Thus, a point $(X,\ Y)$ in the vignetting-corrected image is translated, rotated and scaled to a point $(x,\ y)$ in the preprocessed image according to

$$\begin{bmatrix} x & y & 1 \end{bmatrix}^T = \boldsymbol{H_R H_S H_T} \begin{bmatrix} X & Y & 1 \end{bmatrix}^T, \tag{1}$$

$$\boldsymbol{H}_R = \begin{bmatrix} 1 & r_y & 0 \\ r_x & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ \boldsymbol{H}_S = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \ \boldsymbol{H}_T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ t_x & t_y & 1 \end{bmatrix}. \tag{2}$$

Finally, a preprocessed lenslet image $L(x,\ y)$, which has all the lenslets spaced by a constant multiple of pixel pitch and aligned to an orthogonal integer pixel grid, is generated as shown in Fig. 3(b). The radius of all the macropixels becomes to be m pixels after preprocessing and each macropixel spaces from the neighboring macropixels by h pixels horizontally and v pixels vertically.

The preprocessed lenslet image plays an important role in various applications of plenoptic cameras, such as depth estimation [35], [36], integral displaying [4] and saliency detection [5]. While, novel compression schemes are desired to store and transmit them efficiently. The proposed approach will be described in the next section in detail.

III. Proposed Method

Fig. 4 depicts the overall block-diagram of the proposed lenslet image processing and compression scheme. The proposed scheme mainly consists of following steps (the blocks in gray). Before feeding the preprocessed lenslet image to an encoder, image reshaping consisting of macropixel alignment and interpolation is applied to make the image friendly to compression:

1) *Macropixel Alignment* is applied to the preprocessed lenslet image $L\,(x,\ y)$, which is generated by the steps shown in Fig. 2, to re-organize the macropixels in a way that is aligned with the coding unit grid of a block-based image/video coding standard;
2) *Interpolation* is then employed to generate the intensity value for the missing pixels to maximize the intensity continuity among the regularized adjacent macropixels in the lenslet image; After the reshaped lenslet image is compressed and de-compressed,
3) *Inverse Reshaping* is applied to the decoded image using the decoded camera parameters to convert it back to the lenslet image structure of $L\,(x,\ y)$ as shown in Fig. 3(b). The inverse reshaping basically performs the inverse operations of padding and macropixel alignment.

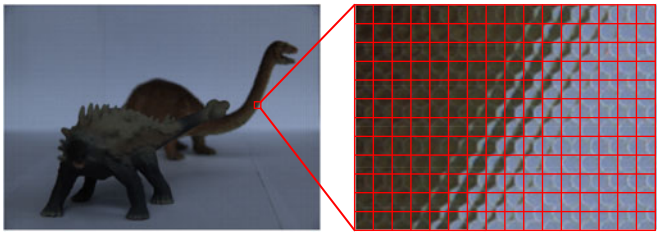


Fig. 5. Partitioning lenslet image by 16×16 coding unit grid.

The proposed image re-shaping and inverse image re-shaping algorithm is described in detail as follows.

A. Macropixel Alignment

Macropixel alignment is to reshape and regularize the lenslet image to make it friendly to block-based coding architecture. Because the adjacent macropixels are spaced one from the other according to the lenslet structure and the bounding box of a macropixel (the block lined in red in Fig. 3(b)) is overlapped with its neighborhoods, directly compressing the lenslet image using a block based video/image codec will result in a coding unit (a non-overlapped block) contains data from multiple macropixels, which is unfriendly to compression. Using the preprocessed lenslet image captured by *Lytro Illum* in Fig. 5 as example, the macropixels in it are spaced one from the other by 16 pixels horizontally and 14 pixels vertically ($h = 16$, $v = 14$ as shown in Fig. 3(b)). If compressing the image by AVC/H.264, the image is decomposed into non-overlapped 16×16 macroblocks, as the red grids shown in Fig. 5, each of which contains data from multiple macropixels with different texture features. This kind of macroblock partitioning is unfriendly to intra-coding tools, such as directional intra prediction, as a macroblock may contain different textures, which makes efficient intra prediction from the neighboring reference samples difficult. Consequently, macropixel alignment is proposed to reshape and regularize the lenslet image to align the macropixel structures with the coding unit partitions.

Considering that the coding unit size used in the existing image/video coding standards is generally an integer multiple of 8 [19], [20], the proposed macropixel alignment reshapes and regularizes the lenslet image to guarantee that the centers of macropixels are aligned and each macropixel can be fully contained by a non-overlapped 16×16 block.

In order to first separate the macropixel rows vertically, a vertical coordinate transformation is proposed to move each macropixel row in $L(x, y)$ to an initially zero valued lenslet image $L_t(x, y)$ according to

$$L_t(x, y + lk) = L(x, y) \quad x, y \in Z, x \in [1, W], y \in [1, H],$$
$$k = \lceil (y - y_0 - 1)/v \rceil, l = n - v, \tag{3}$$

where $L_t(x, y)$ represents the newly generated image after vertical coordinate transformation; (x, y) is the pixel coordinate in the image; W and H are the image width and height of $L(x, y)$; $n \times n$ is coding unit size marked as black box in Fig. 6, where n is set to be 16; v is the vertical spacing between the neighboring two macropixel rows measured by the center distance shown in Fig. 3(b). v is determined by the microlens array structure and can be retrieved for a specific plenoptic camera. l represents the vertical offset used for separating the neighboring two macropixel rows, where $l = n - v$; y_0 represents the vertical spatial offset from the top boundary of the image $L(x, y)$, as shown in Fig. 3(b), which equals to the maximum vertical pixel number in the incomplete macropixel row at the image upper boundary. y_0 is determined by the relative position between the microlens array and the imaging sensor, and varies from one camera to another. It can be derived by preprocessing.

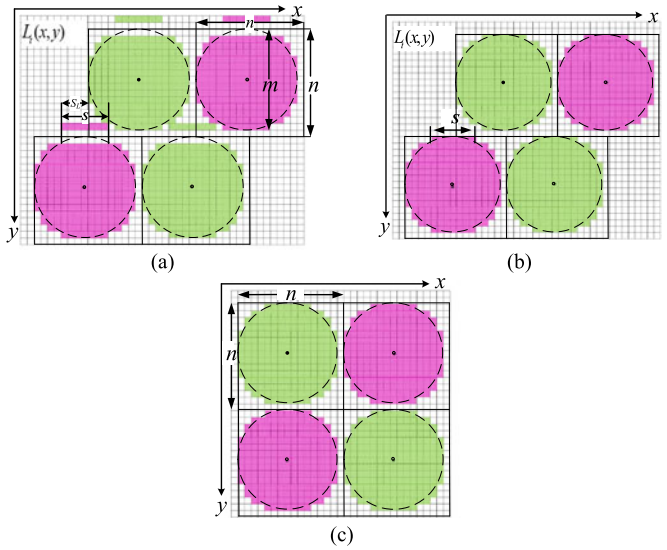


Fig. 6. Macropixel alignment: (a) The image after vertical coordinate transformation; (b) The image after moving S pixels back to the original macropixel; (c) The image after horizontal transformation.

After the vertical coordinate transformation, each macropixel row in Fig. 3(b) is moved downward by l pixel lines, resulting an image with vertical macropixel separation shown in Fig. 6(a).

After the vertical coordinate transformation, some pixels are separated from the original macropixels (the S pixels marked in magenta and green in Fig. 6(a)). Accordingly, it is necessary to move them back to the original macropixels by

$$L_t(pn + x_0 + i, qn + y_0 + l) = L_t(pn + x_0 + i, qn + y_0)$$
$$i, p, q \in Z, i \in [-S_L, S - S_L]$$
$$p \in [0, \lfloor (W_t - x_0)/n \rfloor],$$
$$q \in [0, \lfloor (H_t - y_0)/n \rfloor], \tag{4}$$

where W_t and H_t represent the width and height of $L_t(x, y)$, respectively; p and q represent the horizontal and vertical macropixel coordinate, respectively; S is the number of pixels to be moved back to each macropixel; S_L is the left pixel offset of the to-be-moved pixels relative to the boundary of the coding unit grid, as shown in Fig. 6(a); x_0 is the horizontal spatial offset of each macropixel row from the left boundary of the image, as shown in Fig. 3(b), which equals to the maximum horizontal pixel number in the incomplete macropixel at the image left boundary for the current macropixel row. Generally, x_0 varies between the odd and the even macropixel row and

is a constant to all the odd/even macropixel rows. S_L and x_0 are all achievable based on optical parameters of the plenoptic cameras and preprocessing. The image after this step is shown in Fig. 6(b).

Then, a horizontal coordinate transformation is applied to the odd or the even macropixel rows, whichever have smaller x_0, to align the macropixel centers vertically by shifting macropixel rows to the left by $n/2$ pixels. Selecting to move the macropixel rows with smaller x_0 is based on the consideration that fewer pixels need to be padded afterwards to minimize the image size after macropixel alignment. The aligned image $L_t\,(x,\,y)$ is shown in Fig. 6(c), in which macropixels in $L_t(x,\,y)$ are spaced regularly and fully contained by non-overlapped 16×16 blocks.

After the vertical and horizontal coordinate transformation, boundary processing is applied to the incomplete macropixels along the image boundaries to maintain the coherence between the coding unit block and incomplete macropixels. The process consists of the following five steps.

1) Left image boundary processing: according to the coordinates of macropixels at the boundary, the integrity and pixel width of incomplete macropixels can be verified. For the macropixel rows on which the horizontal coordinate transformation performed, the x_0 incomplete pixel columns on the left boundary are discarded since they are not valid in generating the light field [37]. For the macropixel rows on which the horizontal coordinate transformation is not performed, i.e., those with larger x_0, $(n - x_0)$ pixel columns are padded on the left boundary to complete the block structure.
2) Right image boundary processing: the $n/2$ columns of pixels in the right boundary are checked by verifying whether they are valid pixels for generating light field or not. If they are marked as invalid pixels, they are discarded. Otherwise, pixel padding is performed on the right boundary for the macropixel rows on which the horizontal coordinate transformation is performed by duplicating the right-most pixel column $n/2$ times.
3) Upper image boundary processing: padding is performed on the upper image boundary by duplicating the top pixel row $(m - y_0)$ times to complete the block structure, where m is the radius of each macropixel as depicted in Fig. 3(b). m is determined by the geometry information of microlens and the relative location between the microlens and the imaging sensor. It can be retrieved by preprocessing as described in Section II. Using *Lytro Illum* as an example, m equals to 15 which represents that each microlens covers 15×15 pixels on the sensor.
4) Bottom image boundary processing: pixel validity for generating light field is also checked at the bottom image boundary. If the pixel rows are not valid, they are discarded.
5) Finally, the image width and the image height will be checked to verify whether they are multiple of 8. If not, several pixel rows/columns will be padded at the bottom/right boundary.

Finally, a regularized and reshaped lenslet image is generated consisting of integer number of $n \times n$ coding blocks, each of which contains one macropixel. After the proposed macropixel alignment, the width and the height of the lenslet image becomes to be:

$$\begin{cases} W_t = W + n - x'_0 - n/2 + P_h & f_{\mathrm{HP}} = 0 \\ W_t = W + n - x'_0 + P_h & f_{\mathrm{HP}} = 1 \\ x'_0 = \max\left(x_{0,e}, x_{0,o}\right) \end{cases}$$

$$\begin{cases} H_t = (m - y_0) + \lfloor (H - y_0)/v \rfloor\, l + H + P_v & y_0 > 0 \\ H_t = \lfloor H/v \rfloor\, l + H + P_v & y_0 = 0 \end{cases},$$

$$\begin{cases} P_v = P_v - D_V & f_{\mathrm{VP}} = 0 \\ P_v = P_v & f_{\mathrm{VP}} = 1 \end{cases} \tag{5}$$

where $x_{0,e}$ and $x_{0,o}$ represents x_0 of the even and the odd macropixel row, respectively; f_{HP} and f_{VP} are the flags representing whether the pixels on the right/bottom boundary are valid for light field generation. If not valid, corresponding to $f_{\mathrm{HP}} = 0/f_{\mathrm{VP}} = 0$, $n/2/D_V$ columns/rows of pixels will be removed, where D_V is the number of invalid pixel rows at the bottom image boundary. P_h and P_v represents the number of pixel columns/rows padded on the right/bottom image boundary to guarantee that the image width/height is the multiple of 8, respectively.

The image structure after macropixel alignment is friendly to block based hybrid encoding architecture as each macropixel is fully contained by a coding unit. However, some of pixels in the image after macropixel alignment do not have exact values, such as the white pixels in each block shown in Fig. 6(c). The proposed *Interpolation* is used to determine pixel values for those pixels that do not exist in the original lenslet image or do not valid for light field generation.

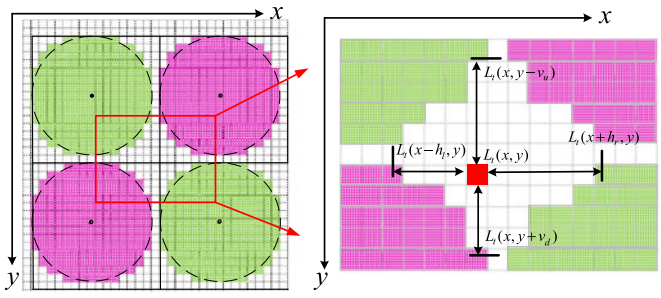


Fig. 7. Proposed interpolation method.

B. Interpolation

After macropixel alignment, the pixels outside the macropixel in each coding block are generated without exact intensity values. In order to maximize the intensity continuity of adjacent macropixels and minimize the possible blocking artifacts, an adaptive pixel interpolation is used to fill those pixels based on the relative distance to the nearest neighborhoods.

As shown in Fig. 7, the white pixels at the four corners and the bottom/right boundary of each coding block need to be filled. Based on their spatial locations, a horizontal edge detector H_{detector} and a vertical edge detector V_{detector}, which represent

the horizontal and vertical gradient of intensity, respectively, are computed. The detectors at a to-be-interpolated pixel $(x,\ y)$ are calculated according to

$$V_{\text{detector}} = (L_t(x, y - v_u) - L_t(x, y + v_d))/(v_u + v_d)$$
$$H_{\text{detector}} = (L_t(x - h_l, y) - L_t(x + h_r, y))/(h_l + h_r), \quad (6)$$

where v_u and v_d, as shown in Fig. 7, represent the distance, measured by the number of pixels, from the pixel to the nearest available pixels above and below the target pixel in the vertical direction, respectively. h_l and h_r represent the distance from the pixel to the nearest left and right available pixels in the horizontal direction, respectively. The horizontal and vertical edge detectors reflect the intensity variations along the horizontal and vertical direction.

Thus, the intensity value of a to-be-interpolated pixel is generated by:

$$L_t(x, y) = w_v(L_t(x, y + v_d) + v_d \times V_{\text{detector}}) + w_h(L_t(x + h_r, y) + h_r \times H_{\text{detector}}), \quad (7)$$

where w_v and w_h represent the vertical and horizontal weights, respectively.

w_v and w_h are adjusted by horizontal and vertical edge detectors. If H_{detector} is greater than V_{detector}, suggesting a lager gradient of intensity changes in horizontal direction, w_v should be larger to keep the continuity among the adjacent pixels. On the contrary, if H_{detector} is smaller than V_{detector}, suggesting a lager gradient of intensity changes in vertical direction, w_h should be lager. The exact values are determined empirically. Finally, a guided filter [38] is performed on the interpolated pixels to further improve the intensity consistency among the adjacent macropixels.

The effectiveness of the proposed interpolation method is demonstrated in [30], in which several interpolation methods were compared. The experimental results showed that the proposed interpolation method outperforms filling the to-be-interpolated pixels by the average value of existing adjacent pixels, filling the pixels by white color (the intensity equals to 255), and filling pixels by black color (the intensity equals to 0) at all compression ratios. It demonstrated that the proposed interpolation method can efficiently improve the intensity consistency among the adjacent macropixels.

The proposed image reshaping is equally applied to the luminance and chrominance components of a lenslet image. After reshaping, the reshaped and regularized lenslet image is fed to a block based video encoder for compression.

C. Imaging Parameters and Inverse Reshaping

After encoding and decoding, the reconstructed lenslet image, which keeps the structure as that after macropixel alignment and interpolation, needs to be converted back to the original macropixel structure for further applications.

In order to perform the inverse image reshaping after decoding, imaging parameters need to be transmitted. Table I summarizes the imaging parameters needed. Setting the maximum m as 32 is big enough as the existing plenoptic cameras normally have much smaller m to guarantee the spatial resolution for the angular information. The parameters are fixed for a specific camera and need to be signaled only once for an image or a sequence. Compared to the overall amount of the bits required for compressing plenoptic content, the overhead consumed by the parameters is negligible.

TABLE I
IMAGING PARAMETERS NEEDED FOR TRANSMISSION

Params.	Semantics	Dynamic Range	Bits
m	Radius of each macropixel	[2,32]	5
v	Vertical spacing between the neighboring two macropixel rows	[2,32]	5
y_0	The maximum vertical pixel number in the incomplete macropixel row at the image upper boundary	[1,31]	5
S	The number of pixels to be moved back to each macropixel	[1,31]	5
S_L	the left pixel offset of the to-be-moved pixels relative to the boundary of the coding unit grid	[1,16]	4
$x_{0,o}$	Incomplete pixels counting from the left boundary of the odd macropixel rows	[2,32]	5
$x_{0,e}$	Incomplete pixels counting from the left boundary of the even macropixel rows	[2,32]	5
P_h	The number of pixel columns padded to the right image boundary	[0,7]	3
P_v	The number of pixel rows padded to the bottom image boundary	[0,7]	3
f_{HP}	Flag representing whether the pixels on the right boundary of the non-transformed macropixel row are valid for light field generation	1, valid 0, else	1
f_{VP}	Flag representing whether the pixels on the bottom boundary of image are valid for light field generation	1, valid 0, else	1

Using the signaled imaging parameters, the reconstructed lenslet image can be converted back to the original macropixel structure by discarding the padded pixels and reversing the macropixel alignment. After the inverse image reshaping, the reconstructed lenslet image $L'(x,\ y)$ maintains exactly the same data structure as the original lenslet image, which can be directly applied to a wide variety of applications.

D. Correlation Analysis in the Reshaped Lenslet Image

After macropixel alignment and interpolation, the macropixels in $L_t(x,\ y)$ are spaced regularly, each of which can be fully

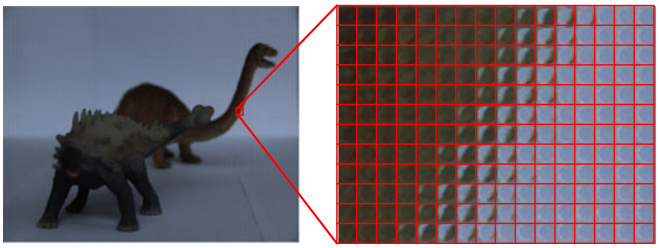


Fig. 8. Partitioning the reshaped lenslet image by 16×16 coding unit grid.

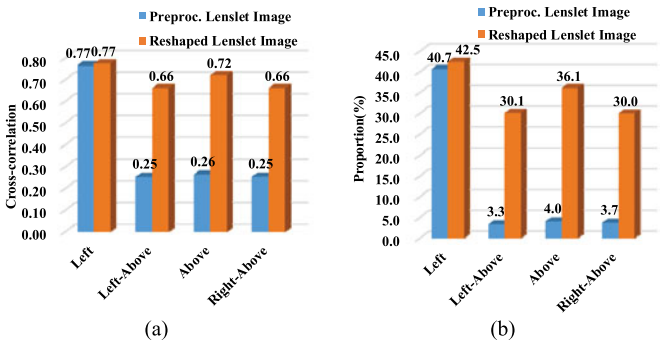


Fig. 9. The statistics of spatial correlations: (a) The average of cross-correlation; (b) The proportion of blocks with cross-correlation higher than 0.9.

contained by a non-overlapped 16×16 block. Like the image shown in Fig. 5, without the proposed image reshaping, the red grid corresponding to 16×16 coding unit grid partitions several incomplete macropixels into one block, which results in low correlations among the adjacent blocks and complex textures inside a block. While, after applying the proposed image reshaping, as shown in Fig. 8, each 16×16 block consists of data only from one macropixel. If a block-based hybrid video encoder compresses the reshaped lenslet image, high spatial correlations among adjacent macropixels can be easily detected and exploited.

To verify this, an analysis is conducted by calculating the average of cross-correlation coefficients between the current 16×16 block and its adjacent four 16×16 blocks located left to, left-above, above and right-above the current block. The correlation statistics both for the preprocessed lenslet image and the reshaped lenslet image are retrieved and shown in Fig. 9 using image "Foutain_&_Vincent" [34] as an instance. It is obvious that after applying the proposed image reshaping, corresponding to the bars in orange, the average of cross-correlation between the current block and its four adjacent blocks is much higher than the preprocessed lenslet image as shown in Fig. 9(a). Notably, the proportion of the blocks with cross-correlation higher than 0.9 in the reshaped image, as shown in Fig. 9(b), is much larger than that in the preprocessed lenslet image also.

It indicates that image reshaping improves the spatial correlations among the spatially adjacent blocks, which will further benefit the spatial coding tools in reducing the spatial redundancy. For example, in the video coding standards like HEVC or H.264/AVC, intra prediction is performed based on the boundary samples in the reconstructed adjacent blocks. Due to the high correlations among adjacent macropixels in the reshaped lenslet image, the directional intra prediction modes rather than DC or planar modes have high potential to be selected to improve the accuracy of intra prediction and to reduce the energy of residual signal. Also, for some new intra prediction tools like displacement intra prediction, image reshaping can provide more precise intra prediction and reduce the overhead introduced by big motion vectors to some extent. The statistical data retrieved for the directional intra prediction modes and the compression performance of working with new intra prediction tools will be provided in Section IV to further demonstrate them.

IV. Experiments and Analysis

In this section, the effectiveness of the proposed lenslet image reshaping method is evaluated using the state-of-the-art video coding standards and new intra prediction tools. First, the test conditions are described. Secondly, the compression efficiency is compare and discussed. Thirdly, further analysis is conducted in terms of intra mode selection, residual energy and computational complexity.

A. Test Conditions

The effectiveness of the proposed scheme for compressing plenoptic content is demonstrated by using the state-of-the-art video coding standards. In the proposed scheme depicted in Fig. 4, we selected three state-of-the-art video coding standards: HEVC/H.265 Format Range Extension (REext) [39], AVC/H.264 High 4:4:4 profile [40] and HEVC/H.265 Screen Content Coding Extension (SCC) [41]. The HEVC reference software HM-16.9_SCM8.0 [42] using the configure file of main-REext, the H.264/AVC reference software JM 19.0 [43] and the HEVC HM-16.9_SCM8.0 using the configure file of main-SCC with all the SCC coding tools turned on are used as reference code for HEVC REext, H.264/AVC and HEVC SCC, respectively. Since the resolution of lenslet image is much higher than the profiles allowed in H.264/AVC, the storage and the reference buffers have been changed in JM 19.0 to support the compression. Five raw lenslet images downloaded from light-field image dataset [34], Fig. 10(a)-(e) captured by *Lytro Illum*, and five raw images captured by our *Lytro Illum* camera, Fig. 10(f)-(j), are tested. The spatial resolution and angular resolution are 7728×5368 and 15×15 (i.e., $m = 15$), respectively. According to the testing method defined in [37], the images are preprocessed as that described in Section II to generate the preprocessed lenslet images. The spatial resolutions of the preprocessed lenslet images are listed in Table II, the second column counting from the left. Since their height and width are not multiple of 8, the bottom/right pixel rows/columns are padded to make the image size of 8656×6080. The padded preprocessed lenslet images in resolution of 8656×6080 are converted from RGB color space to YUV 4:4:4 color space and fed to HEVC and H.264 encoders to generate the anchor.

To test the proposed method, the preprocessed lenslet images without padding are fed to the proposed reshaping scheme. Based on the parameters specified in Table II, the images are reshaped to generate regularized images with spatial resolution of 8656×6944. Before feeding to HEVC and H.264 encoders, the lenslet images are also converted from RGB color space to

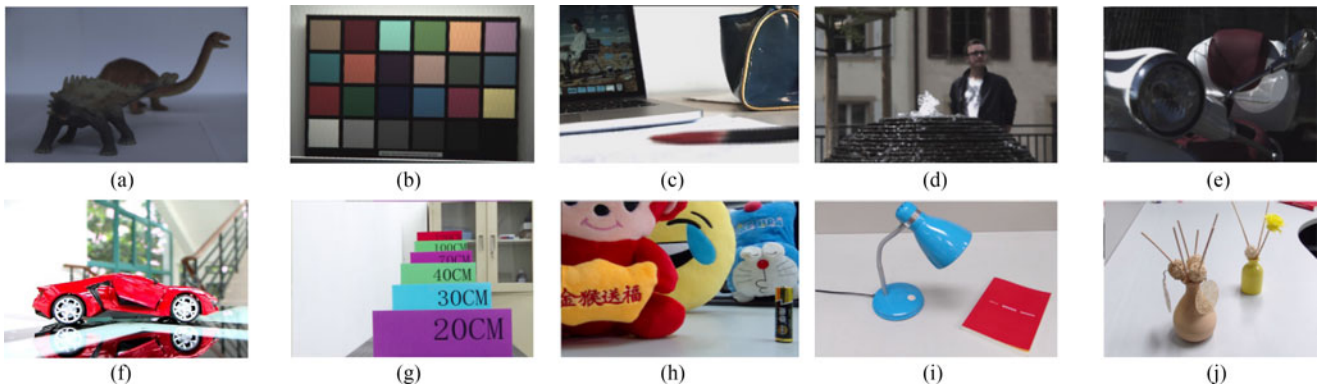


Fig. 10. Testing lenslet images: (a)–(e) are from image dataset [34]; (f)–(j) are images captured by our own *Lytro Illum* camera. (a) Ankylosaurus_&_Diplodocus. (b) Color_Chart_1. (c) Desktop. (d) Foutain_&_Vincent. (e) Vespa. (f) Ferrira_Opendoor. (g) Cards. (h) Dolls. (i) Lamps&Book. (j) Vase.

TABLE II
PARAMETERS OF TESTING IMAGES

Index	Testing Images	Image Size after Preprocessing in Section II	m	v	y_0	S	S_L	$x_{0,o}/x_{0,e}$	Image Size after Reshaping
(a)	Ankylosaurus_&_Diplodocus	8653×6073	15	14	8	7	4	13/5	8656×6944
(b)	Color_Chart_1	8652×6072	15	14	7	7	4	13/5	8656×6944
(c)	Desktop	8655×6074	15	14	9	7	4	14/6	8656×6944
(d)	Foutain_&_Vincent	8654×6073	15	14	8	7	4	14/6	8656×6944
(e)	Vespa	8653×6073	15	14	8	7	4	13/5	8656×6944
(f)	Ferrira_Opendoor	8656×6075	15	14	0	7	4	9/1	8656×6944
(g)	Cards	8656×6074	15	14	0	7	4	9/1	8656×6944
(h)	Dolls	8655×6075	15	14	0	7	4	9/1	8656×6944
(i)	Lamps&Book	8656×6075	15	14	0	7	4	9/1	8656×6944
(j)	Vase	8656×6075	15	14	0	7	4	9/1	8656×6944

TABLE III
CODING PARAMETERS FOR HEVC AND H.264/AVC

Coding Parameters	HEVC REXt Profile	HEVC SCC Profile	H.264 High 4:4:4 Profile
Maximum CU size	64×64	64×64	16×16
Maximum partition depth	4	4	2
QuadtreeTULog2MaxSize	5	5	N/A
QuadtreeTUMaxDepthIntra	3	3	N/A
8×8 Transformation	On	On	On
Loopfilter	On	On	On
SAO	On	On	N/A
Intra Block Copy Enable	N/A	On	N/A
Hash based IBC Enable	N/A	On	N/A
IBC search width in CUs	N/A	3	N/A
IBC nonhash search in CTUs	N/A	1	N/A
RDOQ	On	On	Off
QPs	26, 32, 38, 44		
YUV Format	4:4:4		

YUV 4:4:4 color space. All the images are encoded by "All Intra" configurations. The detailed coding parameters [44] are listed in Table III.

BD-bitrate defined in [45] is used to measure the compression efficiency among the approaches. Compression distortion used in BD-bitrate is calculated between the subaperture images rendered from the original lenslet image and those rendered from the reconstructed lenslet image [37] to reflect the compression effect on generating the light field in the applications. Thus, the BD-bitrate reduction reflects the compression bitrate saving under the same reconstructed light field quality.

B. Experimental Results

1) Intermediate Processing Results of the Proposed Image Reshaping: First, the intermediate processing results of the proposed scheme are illustrated visually in Fig. 11 using "Ankylosaurus_ &_Diplodocus" as an example. Fig. 11(a) to (e) corresponds to the preprocessed lenslet image, the lenslet image after the proposed macropixel alignment, the lenslet image after interpolation, the encoded and reconstructed image and the output image after inverse image re-shaping, respectively. It is obvious that the proposed scheme regularizes the macropixel structure to be compatible with the block-based hybrid coding architecture, and it preserves the valid pixels in the original lenslet image for light field generation to enable lenslet image reconstruction for a wide range of applications. Since the proposed image reshaping is a completely reversible process for the pixels valid in light field generation, it introduces no performance loss or distortion to the light field data. Thus, applying image reshaping before encoding will not introduce additional performance loss to the subperture images rendered from the decoded lenslet images.

2) Comparison Using Different Coding Standards: In order to demonstrate the compatibility and effectiveness of the proposed method to different block-based coding standards, compression efficiency is first compared between the compression results with enabling the proposed image reshaping and those

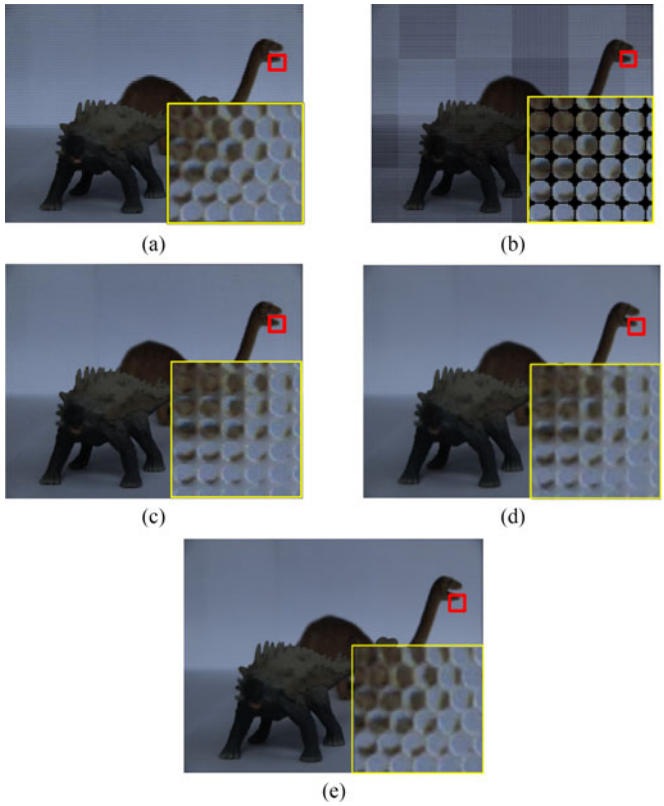


Fig. 11. (a) The lenslet image after preprocessing; (b) The image after macropixel alignment; (c) The image after interpolation; (d) The decoded image at QP = 26; (e) The reconstructed lenslet image after inverse reshaping.

TABLE IV
BD-BITRATE REDUCTION COMPARISON AMONG THE PROPOSED SCHEME, HEVC REXT AND H.264/AVC 444 HIGH PROFILE

Testing Images	HEVC REext w/ Prop. vs. HEVC REext w/o Prop.	H.264/AVC High 444 w/ Prop. vs. H.264/AVC High 444 w/o Prop.
Ankylosaurus_& _Diplodocus	−30.03%	−46.37%
Color_Chart_1	−15.62%	−11.14%
Desktop	−3.50%	−34.11%
Foutain_&_Vincent	−1.50%	−13.34%
Vespa	−1.62%	−34.72%
Ferrira_Opendoor	−1.73%	−14.81%
Cards	−8.61%	−15.47%
Dolls	−12.47%	−38.07%
Lamps&Book	−15.22%	−44.50%
Vase	−14.43%	−23.89%
Average	−10.47%	−27.64%

without enabling the image reshaping using HEVC/H.265 REext and AVC/H.264 High 4:4:4 as coding platform, respectively. Compression efficiency comparison results are summarized in Table IV. As shown in the table, by compressing the reshaped lenslet images, HEVC/H.265 REext and H.264/AVC High 4:4:4 profile can achieve an average of 10.47% and 27.64% bitrate reduction under the same reconstructed light field quality, respectively, compared with compressing the preprocessed lenslet images directly. The peak bitrate reduction achieved for the two different codecs are 30.03% and 46.37%, respectively. Table V lists the absolute bitrate reduction, proportion of bitrate reduction and absolute distortion variation at each QP. It can be observed that compressing the reshaped images can provide 11.47%/33.46% absolute bitrate reduction with only 0.10 dB/0.24 dB degradation in the reconstructed light field quality, which makes a good trade-off between the reconstruction quality and bitrate reduction. The experimental results demonstrate that the reshaped and regularized the macropixel structure in the lenslet images generated by the proposed algorithm greatly benefits the compression using the existing block-based compression standards.

Fig. 12 (a)–(c) show the rate distortion (RD) curves for three randomly selected images. As shown in figures, compressing the reshaped images can outperform the corresponding coding standard without image reshaping at all the testing bitrates. It provides obvious improvement from high compression ratio, corresponding to lower bitrate, to low compression ratio, corresponding to higher bitrate, no matter how efficient the codec is.

3) Comparison Using Different Intra Prediction Tools: To demonstrate that the proposed approach can work together with other intra prediction tools and the inherent correlations in the reshaped lenslet image can be efficiently exploited by the intra prediction methods, the proposed method is further tested with two intra prediction tools Intra-Block Copy (IBC) mode and LLE based 3D holoscopic coding method [11]. IBC mode is a new coding tool employed in HEVC/H.265 SCC Extension [41]. It can be regarded as motion compensation within the current picture. It offers obvious improvement in compression efficiency for the images with repeated patterns and numerically identical blocks, such as lenslet images. LLE mode estimates the current prediction unit by a linear combination of k-nearest spatially neighboring patches. It can bring an obvious compression improvement for holoscopic contents. Five cases as listed in Table VI have been tested.

Table VII compares the compression efficiency among the five cases. It can be found that IBC/LLE modes are effective in compressing the lenslet images. *HEVC + IBC/HEVC + LLE* provides an average of 26.85%/29.64% bitrate reduction by comparing with *HEVC*, respectively. While, more importantly, turning on the proposed algorithm, i.e., compressing the reshaped lenslet images using the two intra prediction tools, can further improve the compression efficiency by an average of 11.62%/9.81% bitrate reduction, as those illustrated in the column of "*Proposed + IBC* vs. *HEVC + IBC*" and "*Proposed + LLE* vs. *HEVC + LLE*". Compared with *HEVC*, the proposed algorithm working with IBC mode and that working with LLE mode, as illustrated in columns of "*Proposed + IBC* vs. *HEVC*" and "*Proposed + LLE* vs. *HEVC*" in Table VII, achieve a maximum of 74.30%/77.56% bitrate reduction with an average of 35.77%/37.67% bitrate reduction. Also, as shown in Fig. 12 (d)-(f), the intra prediction tool compressing the reshaped lenslet images can bring an obvious compression efficiency improvement at most of the testing bitrates. It demonstrates that the proposed algorithm is effective to intra prediction modes like IBC/LLE in further exploiting the inherent correlations among the neighboring macropixels. It brings additional compression

TABLE V
ABSOLUTE COMPRESSION EFFICIENCY COMPARISON

		HEVC REXt w/ Prop. vs. HEVC REXt w/o Prop.			H.264/AVC High 444 w/ Prop. vs. H.264/AVC High 444 w/o Prop.		
Testing Images	QP	△Bitrate (kbps)	Percentage of △Bitrate	△PSNR (dB)	△Bitrate (kbps)	Percentage of △Bitrate	△PSNR (dB)
Ankylosaurus_&_Diplodocus	26	−4066.576	−20.94%	−0.21	−3663.512	−15.67%	−0.43
	32	−1156.096	−23.14%	0.24	−5488.792	−40.32%	−0.16
	38	−99.856	−11.82%	0.11	−4471.600	−64.69%	−0.09
	44	−7.576	−3.24%	0.14	−3783.288	−59.78%	−0.40
Color_Chart_1	26	−2763.792	−9.91%	−0.30	−2960.336	−8.52%	−0.55
	32	−1163.872	−12.04%	−0.03	371.144	2.53%	−0.45
	38	−606.136	−18.63%	0.12	−4384.024	−40.73%	−0.26
	44	−162.464	−17.09%	0.14	−4288.664	−56.10%	−0.36
Desktop	26	−384.992	−2.36%	−0.27	−700.384	−3.34%	−0.26
	32	−115.656	−1.78%	−0.09	−4274.528	−30.50%	−0.17
	38	−199.728	−7.28%	−0.03	−4322.736	−46.26%	−0.14
	44	−124.952	−11.44%	0.07	−4183.536	−56.50%	−0.33
Foutain_&_Vincent	26	1609.992	5.10%	−0.27	1537.368	3.86%	−0.31
	32	−363.064	−2.51%	−0.20	−480.680	−2.41%	−0.24
	38	−680.448	−11.02%	−0.14	−4903.368	−34.34%	−0.13
	44	−353.240	−16.12%	0.07	−4843.264	−54.16%	0.03
Vespa	26	−206.816	−1.32%	−0.15	−454.280	−2.15%	−0.32
	32	−302.504	−4.86%	−0.11	−4233.936	−30.18%	−0.20
	38	−112.128	−4.89%	−0.03	−4179.864	−48.15%	−0.09
	44	−73.600	−9.20%	−0.01	−4044.384	−60.02%	−0.18
Ferrira_Opendoor	26	854.928	2.69%	−0.31	265.696	0.67%	−0.32
	32	−453.800	−3.15%	−0.17	−918.552	−4.65%	−0.25
	38	−437.272	−6.95%	−0.01	−4753.200	−34.09%	−0.08
	44	−214.488	−8.42%	0.04	−4508.912	−48.04%	−0.10
Cards	26	−4114.84	−8.34%	−0.53	−4860.768	−8.57%	−0.58
	32	−3021.864	−19.51%	−0.32	−3392.712	−14.76%	−0.36
	38	−1015.792	−26.59%	−0.06	−5192.176	−42.07%	−0.15
	44	−224.616	−21.05%	0.03	−4374.168	−55.40%	−0.07
Dolls	26	−4303.264	−17.23%	−0.43	−4010.504	−13.05%	−0.34
	32	−1199.104	−16.25%	−0.08	−5386.744	−33.71%	−0.18
	38	−300.912	−12.63%	0.04	−4459.392	−50.51%	−0.04
	44	−113.840	−13.85%	0.00	−4074.128	−56.87%	−0.22
Lamps&Book	26	−8647.912	−26.88%	−0.58	−8083.400	−22.43%	−0.47
	32	−1861.544	−34.12%	−0.07	−6606.792	−44.81%	−0.16
	38	−6.680	−0.84%	0.03	−4203.44	−60.60%	−0.11
	44	26.008	9.71%	0.04	−3965.072	−61.08%	−0.05
Vase	26	−5971.968	−11.49%	−0.57	−6911.464	−11.62%	−0.55
	32	−4745.456	−30.09%	−0.34	−4843.056	−22.29%	−0.29
	38	−946.888	−29.10%	0.09	−5504.952	−50.40%	−0.03
	44	−2.408	−0.33%	0.01	−4047.032	−56.68%	−0.14
Average		−1200.880	−11.47%	−0.10	−3839.586	−33.46%	−0.24

efficiency improvements to intra prediction tools as they compress the reshaped lenslet images, which is beneficial to the most advanced video codecs such as HEVC SCC.

C. Analysis of the Proposed Method

To further analyze the compression efficiency improvement, the statistics of coding units (CUs) on different quadtree depths in HEVC is retrieved by using "Ankylosaurus_&_Diplodocus" as an instance. First, using 4×4 block as the basic unit, Table VIII compares the proportion of CU on different quadtree depths between HEVC REXt with and without the proposed method for the luminance component. As shown in the table, compared with directly encoding the preprocessed lenslet images, smaller CU sizes, corresponding to deeper partition depth that can provide more precise prediction and avoid ringing artifacts, are more likely to be selected by the encoder when the lenslet images are reshaped by the proposed approach. Then, still using 4×4 block as the basic unit, statistical data are retrieved for Intra prediction directions for coding luminance component to further illustrate the efficiency of the proposed encoding scheme. As shown in Table IX, when directly compressing

TABLE VI
CODING CONFIGURATIONS OF TESTING CASES

Testing cases	Configurations
HEVC	HEVC SCC Profile with the settings in Table III except that IBC [41] is turned off and the proposed algorithm (the gray blocks in Fig. 4) is bypassed
HEVC + IBC	"HEVC" with IBC on
Proposed + IBC	"HEVC + IBC" + the proposed image reshaping algorithm
HEVC + LLE	"HEVC" + LLE mode [11]
Proposed + LLE	"HEVC + LLE" + the proposed image reshaping algorithm

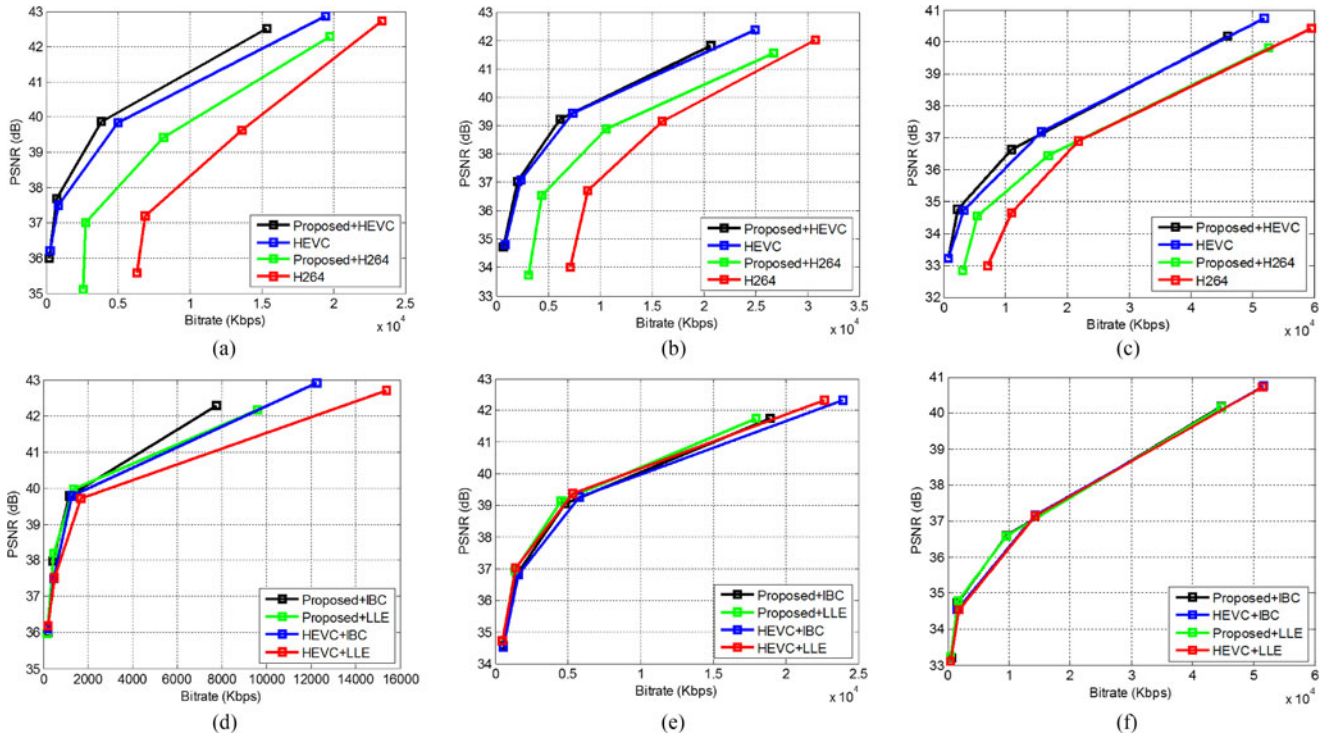


Fig. 12. Rate distortion curves for light field compression efficiency: Different coding standards: (a) Ankylosaurus_&_Diplodocus; (b) Dolls; (c) Vase; Different intra prediction modes: (d) Ankylosaurus_&_Diplodocus; (e) Dolls; (f) Vase.

TABLE VII
BD-BITRATE COMPARISON AMONG THE FIVE TESTING CASES IN TABLE VI

Testing Images	HEVC+IBC vs. HEVC	Proposed+IBC vs. HEVC+IBC	Proposed+IBC vs. HEVC	HEVC+LLE vs. HEVC	Proposed+LLE vs. HEVC+LLE	Proposed+LLE vs. HEVC
Ankylosaurus_& _Diplodocus	−45.60%	−26.1%	−61.10%	−43.65%	−30.86%	−60.91%
Color_Chart_1	−66.20%	−19.30%	−74.30%	−74.48%	−6.74%	−77.56%
Desktop	−24.10%	−4.30%	−27.80%	−33.10%	3.40%	−32.10%
Foutain_&_Vincent	−32.20%	−2.60%	−34.40%	−41.20%	4.80%	−39.21%
Desktop	−23.50%	−6.50%	−28.80%	−21.44%	−8.65%	−28.41%
Ferrira_Opendoor	−25.50%	−1.00%	−26.40%	−18.61%	−22.12%	−36.69%
Cards	−19.80%	−12.40%	−31.00%	−30.00%	3.70%	−27.40%
Dolls	−13.40%	−13.80%	−26.20%	−20.89%	−9.29%	−29.61%
Lamps&Book	−4.80%	−17.60%	−22.00%	−1.22%	−17.75%	−19.05%
Vase	−13.40%	−12.60%	−25.70%	−11.81%	−14.61%	−25.77%
Average	−26.85%	−11.62%	−35.77%	−29.64%	−9.81%	−37.67%

TABLE VIII
PROPORTION OF CODING UNITS (CU) ON DIFFERENT QUADTREE DEPTHS

	HEVC RExt without the proposed method				HEVC RExt with the proposed method			
QP	Depth 0	Depth 1	Depth 2	Depth 3	Depth 0	Depth 1	Depth 2	Depth 3
26	72.40%	24.29%	1.70%	1.62%	46.31%	47.35%	2.41%	3.93%
32	77.04%	21.43%	1.21%	0.32%	73.62%	24.67%	1.38%	0.33%
38	79.91%	18.96%	0.95%	0.17%	86.65%	12.47%	0.73%	0.15%
44	92.50%	6.86%	0.62%	0.03%	87.17%	12.16%	0.60%	0.06%

TABLE IX
PROPORTION OF DIRECTIONAL PREDICTION MODES FOR INTRA PREDICTION

	HEVC REXt without the proposed method					HEVC REXt with the proposed method				
QP	0:Planar	1:DC	10: Horizontal	26: Vertical	Others	0:Planar	1:DC	10: Horizontal	26: Vertical	Others
26	7.99%	83.99%	1.97%	0.62%	5.44%	15.05%	52.94%	4.75%	9.29%	17.98%
32	8.66%	75.16%	8.80%	2.15%	5.23%	9.74%	20.60%	20.90%	42.92%	5.84%
38	18.74%	32.26%	28.79%	10.38%	9.83%	11.29%	10.44%	30.14%	43.17%	4.97%
44	42.11%	14.89%	7.13%	16.00%	19.87%	41.06%	14.23%	5.63%	23.38%	15.69%

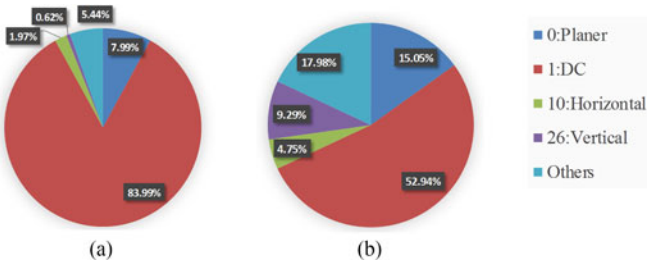


Fig. 13. Proportions of directional intra prediction at QP26: (a) HEVC REXt w/o the proposed approach; (b) HEVC REXt w/ the proposed approach.

the preprocessed lenslet image by HEVC encoder, most of the blocks will select DC prediction mode, corresponding to Intra prediction mode 1, and planar prediction mode, corresponding to Intra prediction mode 0 [20]. While, as compressing the reshaped lenslet images generated by the proposed approach, a much larger portion of blocks will select vertical, horizontal or even more complex directional prediction modes, corresponding to Intra prediction mode 26, 10 and others, respectively. The percentage of using DC prediction mode is greatly reduced. The statistics revealed that the image reshaping greatly improves data correlation for Intra prediction, thus provides better compression efficiency for the existing block-based video coding standards.

Fig. 13 further compares proportion of Intra prediction directions under QP $= 26$ visually. In contrast to directly coding the preprocessed image by HEVC, which mainly selects DC prediction mode, the proposed reshaping approach lets around half of the blocks select planar and directional prediction modes. Although the mode index of DC prediction is simpler than other directional prediction modes, which may consume fewer bits, the prediction result generated by it is always too coarse, which generally results in much larger prediction residue. Fig. 14 shows the residual image generated by the HEVC REXt with and without the proposed method. It is obvious that the energy in the residual image generated by HEVC REXt with the proposed approach is much lower than that without using the proposed approach, which is consistent with the statistical analysis on the partition depth and prediction directions provided above.

As we analyze the compression efficiency improvement among the images, it can be observed that the efficiency improvement of the proposed approach is content dependent. When the proposed image reshaping is applied to the lenslet images, the intensity continuity between the adjacent macropixels is generally improved, but the improvement depends on how complex the texture and color of the images is. As shown in Table IV, for the images with colorful textures, like

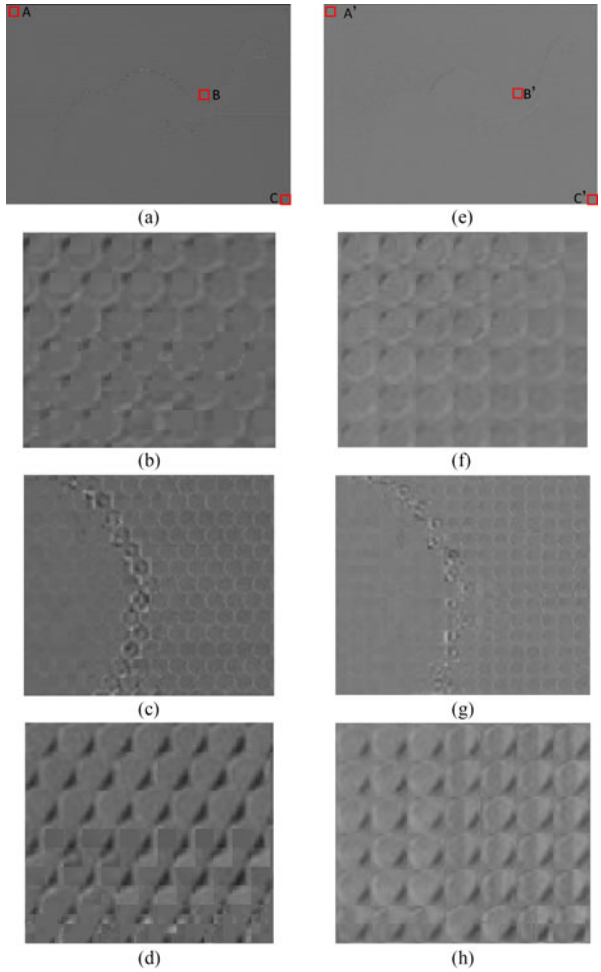


Fig. 14. Residual image of "Ankylosaurus_&_Diplodocus" generated by: (a) HEVC REXt without the proposed method; (e) HEVC REXt with the proposed method; (b)-(d) magnification of A, B and C in (a); (f)-(h) magnification of A', B' and C' on (e).

"Foutain_&_Vincent", lower improvement is achieved in the compression efficiency, while for the images with homogeneous background and simple texture, like "Ankylosaurus_& _Diplodocus", a much higher improvement can be obtained. This content dependency could be leveraged by e.g., designing content adaptable interpolation filters based on smoothness optimization, which is part of our future work.

The section is concluded by analyzing the complexity overhead introduced by the proposed algorithm for encoding and decoding. During macropixel alignment, the major computations are consumed by linear coordinate transformation, which

TABLE X
COMPUTATIONAL COMPLEXITY ANALYSIS

Testing Images	HEVC					H.264/AVC				
	$T_{\text{w/ oIR}}$ (s)	T_{IR} (s)	$T_{\text{E+ IR}}$ (s)	$T_{\text{IR}}/T_{\text{w/ oIR}}$	$(T_{\text{E+ IR}}-T_{\text{w/ oIR}})/T_{\text{w/ oIR}}$	$T_{\text{w/ oIR}}$ (s)	T_{IR} (s)	$T_{\text{E+ IR}}$ (s)	$T_{\text{IR}}/T_{\text{w/ oIR}}$	$(T_{\text{E+ IR}}-T_{\text{w/ oIR}})/T_{\text{w/ oIR}}$
(a)	325.30	0.81	374.97	0.25%	15.27%	143.65	0.79	173.97	0.55%	21.11%
(b)	341.90	0.82	385.76	0.24%	12.83%	166.65	0.82	193.64	0.49%	16.20%
(c)	320.10	0.80	362.96	0.25%	13.39%	111.81	0.84	131.12	0.75%	17.27%
(d)	357.87	0.82	407.90	0.23%	13.98%	120.35	0.79	142.01	0.66%	18.00%
(e)	321.90	0.80	367.58	0.25%	14.19%	144.17	0.81	168.21	0.56%	16.67%
(f)	369.15	0.85	414.23	0.23%	12.21%	165.92	1.14	191.99	0.69%	15.71%
(g)	320.93	0.80	363.71	0.25%	13.33%	112.85	0.73	130.26	0.65%	15.43%
(h)	349.68	0.84	399.16	0.24%	14.15%	118.26	0.59	139.70	0.50%	18.13%
(i)	338.32	0.81	389.94	0.24%	15.26%	158.62	0.82	186.72	0.52%	17.72%
(j)	371.35	0.85	424.38	0.23%	14.28%	157.91	0.81	183.62	0.51%	16.28%
Average	341.65	0.82	389.10	0.24%	13.89%	140.02	0.82	164.17	0.59%	17.25%

involves 7 additions/subtractions, 2 multiplications/ divisions and 1 clipping process for each pixel located apart from the image boundaries. While, the range of i, p and q and the value of $pn + x_0$ and $qn + y_0$ only need to be calculated once for an image or a macropixel row. During interpolation, generating a to-be-interpolated pixel requires 6 multiplications/divisions and 13 additions/subtractions. Since the to-be-interpolated pixels are independent during processing, they can be generated in parallel. The bitstream overhead introduced by the proposed algorithm is the reshaping parameters listed Table I. Only 42 bits are needed which is negligible to the decoder.

Applying image reshaping, the reshaped lenslet images will be 14.2% larger on average, as that shown in Table II, which will introduce a slightly increase in the encoding and decoding complexity. To evaluate the computational complexity, the proposed method is integrated into HEVC reference software HM-16.9_SCM8.0 [42] using the configure file of main-RExt and H.264/AVC reference software JM 19.0 [43]. Using the execution time of encoding the lenslet images without reshaping as the anchor, denoted by $T_{\text{w/ oIR}}$, the execution time consumed by image reshaping, denoted by, and the total execution time consumed by reshaping and encoding, denoted by $T_{\text{E+ IR}}$, are retrieved using a PC with Intel Core™ i7-4790 CPU @ 3.60 GHz with 12 GB RAM and 64-bits Windows 7 operating system. The comparison results including the absolute execution time and the relative ratios are shown in Table X. As shown in the table, the proposed image reshaping only consumes 0.24%/0.59% computational complexity, shown by $T_{\text{IR}}/T_{\text{w/ oIR}}$, for HEVC and H.264/AVC encoding, which is negligible to the encoder. Although the total encoding time increases 13.89%/17.25% on average due to the image size increase in the encoder using the proposed approach, shown by $(T_{\text{E+ IR}} - T_{\text{w/ oIR}})/T_{\text{w/ oIR}}$, 10.47%/27.64% bitrate reduction can be achieved, as shown in Table IV, which provides a good tradeoff between the computational complexity and compression efficiency.

V. CONCLUSION

This paper proposes a novel lenslet image reshaping scheme, which efficiently exploits the correlation among macropixels when compress the lenslet images using the existing block based coding standards. The core part of the proposed scheme is to employ the macropixel alignment and adaptive interpolation to reshape and regularize the lenslet image to be friendly to the block-based coding structure. Applying the proposed scheme to the state-of-the-art block-based video coding standards, i.e., HEVC/H.265 RExt and AVC/H.264, leads to a significant bitrate reduction. A peak bitrate reduction of 30.03% and 46.37% and an average bitrate reduction of 10.47% and 27.64% can be achieved under the same reconstructed light field quality. For some advanced coding tools of lenslet image compression, like IBC/LLE, applying the proposed method provides a further bit-rate reduction of 11.62%/9.81% on average. The proposed scheme presents high compatibility with existing block-based coding standards, which can be of interest to broaden the applications of light field through more efficient storage and transmission. The performance of the proposed scheme could be further improved by designing content adaptive interpolation filters together with specific Intra prediction modes, which are under investigation.

REFERENCES

[1] Lytro. [Online]. Available: https://www.lytro.com/. Accessed on: 2017.
[2] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, Oct. 2017.
[3] M. Martínez-Corral, A. Dorado, H. Navarro, A. Llavador, G. Saavedra, and B. Javidi, "From the plenoptic camera to the flat integral-imaging display," *Proc. SPIE*, vol. 9117, pp. 91170H-1–91170H-6, 2014.
[4] X. Xiao, B. Javidi, M. Martinez-Corral, and A. Stern, "Advances in three-dimensional integral imaging: sensing, display, and applications," *Appl. Opt.*, vol. 52, no. 4, pp. 546–560, 2013.
[5] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Comput. Vis. Pattern Recog.*, 2014, pp. 2806–2813.
[6] A. Ghasemi and M. Vetterli, "Scale-invariant representation of light field images for object recognition and tracking," *Proc. SPIE*, vol. 9020, pp. 159–164, 2014.
[7] O. Johannsen, A. Sulc, and B. Goldluecke, "On linear structure from motion for light field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 720–728.
[8] R. S. Higa, R. F. L. Chavez, and R. B. Leite, "Plenoptic image compression comparison between JPEG, JPEG2000 and SPITH," *Cyber J., J. Sel. Areas Telecommun.*, vol. 3, no. 6, pp. 1–6, 2013.
[9] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Efficient intra prediction scheme for light field image compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, 2014, pp. 539–543.
[10] Y. Li, M. Sjostrom, R. Olsson, and U. Jennehag, "Coding of focused plenoptic contents by displacement intra prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 7, pp. 1308–1319, Jul. 2016.

[11] L. F. R. Lucas *et al.*, "Locally linear embedding-based prediction for 3D holoscopic image coding using HEVC[C]," in *Proc. 2014 22nd Eur. IEEE Signal Process. Conf.*, 2014, pp. 11–15.
[12] Y. Li, M. Sjostrom, and R. Olsson, "Coding of plenoptic images by using a sparse set and disparities," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2015, pp. 1–6.
[13] D. Liu *et al.*, "Disparity compensation based 3D holoscopic image coding using HEVC," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, 2015, pp. 201–205.
[14] A. Aggoun, "Compression of 3D integral images using 3D wavelet transform," *J. Display Technol.*, vol. 7, no. 11, pp. 586–592, 2011.
[15] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections," *ACM Trans. Graph.*, vol. 32, no. 4, 2013, Art. no. 46.
[16] P. Ramanathan and B. Girod, "Rate-distortion analysis for light field coding and streaming," *Signal Process., Image Commun.*, vol. 21, pp. 462–275, Mar. 2006.
[17] F. Dai, J. Zhang, Y. Ma, and Y. Zhang, "Lenslet image compression scheme based on subaperture images streaming," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 4733–4737.
[18] T. Sakamoto, K. Kodama, and T. Hamamoto, "A study on efficient compression of multi-focus images for dense light-field reconstruction," in *Proc. IEEE Visual Commun. Image Process.*, San Diego, CA, USA, 2012, pp. 1–6.
[19] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.
[20] G. J. Sullivan, J. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
[21] D. Liu *et al.*, "Pseudo-sequence-based light field image compression," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops, IEEE Comput. Soc.*, 2016, pp. 1–4.
[22] S. Shi, P. Gioia, and G. Madec, "Efficient compression method for integral images using multi-view video coding," in *Proc. 18th IEEE Int. Conf. Image Process.*, 2011, pp. 137–140.
[23] C. Conti, L. D. Soares, and P. Nunes, "HEVC-based 3D holoscopic video coding using self-similarity compensated prediction," *Signal Process., Image Commun.*, vol. 42, pp. 59–78, 2016.
[24] C. Conti *et al.*, "Light-field video coding using geometry-based disparity compensation," in *Proc. IEEE 2014 Conf. True Vis.-Capture, Transm. Display 3D Video*, 2014, pp. 1–4.
[25] D. Lelescu and F. Bossen, "Representation and coding of light field data," *Graph. Models*, vol. 66, no. 4, pp. 203–225, Jul. 2004.
[26] J. Chen, J. Hou, and L. P. Chau, "Light field compression with disparity guided sparse coding based on structural key views," arXiv:1610.03684v1 [cs.CV], Oct. 12, 2016.
[27] C. L. Chang, X. Zhu, P. Ramanathan, and B. Girod, "Light field compression using disparity-compensated lifting and shape adaptation," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 793–806, Apr. 2006.
[28] C. L. Chang and X. Zhu, "Light field compression using disparity-compensated wavelet decomposition," *IEEE Trans. Image Process.*, vol. 15, no. 4, pp. 793–806, Apr. 2006.
[29] D. Xu, Q. Dai, and W. Xu, "Data compression of light field using wavelet packet," in *Proc. 2004 IEEE Int. Conf. Multimedia Expo*, 2004, pp. 1071–1074.
[30] H. Han, X. Jin, and Q. Dai, "Plenoptic image compression based on linear transformation and interpolation," in *Proc. Asia-Pac. Signal Inf. Process. Assoc. Annu. Summit Conf.*, Jeju, South Korea, 2016, pp. 1–6.
[31] E. Y. Lam, "Computational photography with plenoptic camera and light field capture: Tutorial," *J. Opt. Soc. Amer. A*, vol. 32, no. 11, pp. 2021–2032, 2015.
[32] Lytro Illum. [Online]. Available: https://illum.lytro.com/zh/illum. Accessed on: 2017.
[33] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenslet-based plenoptic cameras," in *Proc. 2013 IEEE Conf. Comput. Vis. Pattern Recog.*, Portland, OR, USA, 2013, pp. 1027–1034.
[34] *Plenoptic Images Dataset*. [Online]. Available: http://mmspg.epfl.ch/EPFL-light- field-image-dataset. Accessed on: 2017.
[35] H. Lin, C. Chen, S. B. Kang, and J. Yu, "Depth recovery from light field using focal stack symmetry," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3451–3459.
[36] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 673–680.
[37] *Light Field Compression Evaluation*. [Online]. Available: http://mmspg.epfl.ch /files/content/sites/mmspl/files/shared/LF-GC/C FP.pdf. Accessed on: 2017.
[38] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.
[39] *HEVC Format Range Extension*. [Online]. Available: http://www.hevc.info/rext. Accessed on: 2017.
[40] *Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification, Document 14 496-10 AVC*, ITU-T Rec. H.264/ISO/IEC, Mar. 2013.
[41] J. Xu, R. Joshi, and R. A. Cohen, "Overview of the emerging HEVC screen content coding extension," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 50–62, Jan. 2016.
[42] [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.9 + SCM-8.0/. Accessed on: 2017.
[43] [Online]. Available: http://iphome.hhi.de/suehring/tml/download/. Accessed on: 2017.
[44] O. C. Au, X. Zhang, C. Pang, and X. Wen, "Suggested common test conditions and software reference configurations for screen content coding," Joint Collaborative Team on Video Coding, Torino, Italy, JCTVC-F696, Jul. 2011.
[45] G. Bjontegaard, *Calculation of Average PSNR Difference Between RD-Curves*, ITU-T VCEG-M33, 2001.



Xin Jin (S'03–M'09–SM'11) received the M.S. degree in communication and information system and the Ph.D. degree in information and communication engineering from Huazhong University of Science and Technology, Wuhan, China, in 2002 and 2005, respectively.

From 2004 to 2005, she was an Intern with the Internet Multimedia Group, Microsoft Research Asia, Beijing, China. From 2006 to 2008, she was a Postdoctoral Fellow with The Chinese University of Hong Kong. From 2008 to 2012, she was a Visiting Lecturer with Information Technology Research Organization, Waseda University, Fukouoka, Japan. Since March 2012, she has been with the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China, where she is currently an Associate Professor. She has published more than 115 conference and journal papers. Her current research interests include power-constrained video processing and compression, computational imaging, and multimedia cloud computing.

Dr. Jin is a Member of the SPIE and the ACM. She is the Chair of 3-D video compression standard ad hoc group of the Audio Video Standard Workgroup of China (AVS). She has served on many conference committees, e.g., the PCM, the VCIP, etc., and served as a reviewer for many transactions, e.g., the IEEE TIP, the TCSVT, etc., and international conferences, e.g., the ICIP, the ISCAS, the ICME, etc. She received the AVS Outstanding Contributor Award of year 2004 and the ISOCC Best Paper Award in 2010.



Haixu Han (S'16) received the B.S. degree from the Department of Automation, Dalian University of Technology, Liaoning, China, in 2015. He is currently working toward the M.E. degree in the Department of Automation, Tsinghua University, Beijing, China.

His current research interests include light-field compression and image/video coding. He has published several conference papers, e.g., the ICME, the ICIP, the APSIPA, etc.



Qionghai Dai (SM'00) received the B.S. degree in mathematics from Shaanxi Normal University, Xi'an, China, in 1987, and the M.E. and Ph.D. degrees in computer science and automation from Northeastern University, Shenyang, China, in 1994 and 1996, respectively. Since being a Postdoctoral Researcher in the Automation Department, he has been with the Media Lab, Tsinghua University, Shenzhen, China, where he is currently an Associate Professor and Head of the Lab. His research interests include signal processing, broadband networks, video processing, and communication.

Superrays for Efficient Light Field Processing

Matthieu Hog, Neus Sabater, and Christine Guillemot, *Fellow, IEEE*

***Abstract*—Light field acquisition devices allow capturing scenes with unmatched postprocessing possibilities. However, the huge amount of high-dimensional data poses challenging problems to light field processing in interactive time. In order to enable light field processing with a tractable complexity, in this paper, we address the problem of light field oversegmentation. We introduce the concept of superray, which is a grouping of rays within *and* across views, as a key component of a light field processing pipeline. The proposed approach is simple, fast, accurate, easily parallelisable, and does not need a dense depth estimation. We demonstrate experimentally the efficiency of the proposed approach on real and synthetic datasets, for sparsely and densely sampled light fields. As superrays capture a coarse scene geometry information, we also present how they can be used for real-time light field segmentation and correcting refocusing angular aliasing.**

***Index Terms*—Image processing, image segmentation.**

I. Introduction

OVER the past few years, light field capturing devices have gained in popularity, both at the research and the industrial level. These devices can be broadly classified into the following categories. Plenoptic cameras use an array of microlenses placed in front of the sensor of a classical camera in order to capture light fields [1], leading to trade spatial resolution for angular resolution. Although this trade-off can be controlled [2], the captured light field is always angularly dense (the disparity is of the order of the pixel). Plenoptic cameras typically target consumer photography via their refocusing feature (e.g. Lytro Illum[1]), but also the industrial market thanks to the capabilities they offer for passive, accurate, monocular depth estimation (e.g. Raytrix[2]). Light fields can also be captured by camera arrays [3], [4]. While being significantly more expensive and harder to use, hence dedicated to professionals, they offer the best spatial resolution (but low angular resolution) and a much larger baseline, making them suitable for applications requiring long range depth estimation, change of viewpoint and view synthesis, such as AR content capture (e.g. Lytro Immerge[3]) or movie post production (e.g. Light-field Technology at Fraunhofer [5][4]). A third category of acquisition devices is the camera gantry (e.g. the Stanford Lego Gantry[5]), where a camera is moved along a plane and shots are taken at regular intervals. These devices are easy to use and implement, and can in theory provide high spatial and angular resolution. However, this type of device is limited to the capture of static light fields. Finally, recent and emerging smart phones are equipped either with several cameras (e.g Iphone 7, Amazon Fire Phone), or with a single specialised sensor (e.g Pelican Imaging [5], ProFUSION-25C [5], or with a Wafer-level-optics camera array [6]) which could, to some extent, capture light fields, not as angularly dense as plenoptic cameras. Because they need to be compact, they also have a much smaller baseline and resolution than camera arrays.

The volume of data inherent to light fields, for all capturing devices, is a real issue for user interaction that requires near real-time processing, potentially on devices having limited computational power. This becomes even more critical for light field videos. A second issue is the ease of use. While users are familiar with 2D image editing, light field editing boils down to edit a 3D capture embedded into a 4D signal [7]. One objective of the paper is to enable user interaction with the whole light field while entering inputs on one view only. State of the art light field editing methods either only deal with densely sampled light fields or use a dense depth estimation to perform user-guided segmentation [8]–[10] or to propagate user inputs [11]–[15]. Despite the latest advances [16]–[19] in light field depth estimation, these methods use computationally expensive regularisation to obtain satisfactory depth maps.

Our goal is instead to propose a solution for unsupervised light field over-segmentation which would in addition be angular sampling agnostic and rely less on depth estimation. Our approach is motivated by the observation that, for most editing applications, it might be more important to have accurate object boundaries, even with coarse depth information, than having a refined depth map. We show in Section IV, as already noticed by the authors in [7], that a dense and accurate (in terms of disparity values) depth estimation is not needed for some typical light field applications.

To treat the aforementioned problems, we introduce in this paper the concept of *super-ray* which is the counterpart of super-pixels [20] for light fields. The major difference with conventional super-pixels and super-voxels is that super-rays group perceptually similar *and* corresponding pixels across several views. In other words, super-rays are groups of rays of similar color coming from the same scene area. We then propose what

Manuscript received December 1, 2016; revised April 10, 2017 and June 28, 2017; accepted July 29, 2017. Date of publication August 11, 2017; date of current version October 23, 2017. The guest editor coordinating the review of this paper and approving it for publication was Prof. Y. Liu. *(Corresponding author: Matthieu Hog.)*

M. Hog and C. Guillemot are with the F337 VERT, INRIA, Rennes 35042, France (e-mail: matthieu.hog@technicolor.com; christine.guillemot@inria.fr).

N. Sabater is with Technicolor, Rennes 35700, France (e-mail: neus.sabater@technicolor.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2738619

[1]http://www.lytro.com/imaging

[2]http://www.raytrix.de

[3]http://www.lytro.com/immerge

[4]http://www.iis.fraunhofer.de/en/ff/bsy/tech/lichtfeld.html

[5]http://lightfield.stanford.edu/lfs.html

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

we believe to be the first light field over-segmentation algorithm. It is inspired by SLIC [21], a state of the art super-pixel generation method with good properties in terms of accuracy over complexity and parallelism. The major difference is that centroids, initialised on a reference view, are shared by all the views with a unique spatial and color coordinate and an attached depth information. Then, the assignment step is simultaneously performed by projecting the centroids onto each view and the update is done by re-projecting all the rays assigned to the super-ray onto the reference view, using a fronto planar assumption. The approach is fast, free of any strong scene geometry prior, easy to understand and to implement, and it gives satisfactory results. Moreover, it does not require a dense depth map estimation, making it a suitable candidate for a first step of a light field processing pipeline.

A new metric is then introduced to evaluate an important feature of super-rays: the view-consistency of the over-segmentation. We also quantitatively and qualitatively test our approach on synthetic and real light field data-sets having different angular and spatial resolutions, using standard super-pixel metrics. Finally, in [7], [12] the authors observe that an accurate dense depth map is not crucial for editing. As a follow-up of this work, we show that super-rays enable light field processing tasks which are typically done using a dense depth map, such as light field segmentation and correcting the angular aliasing that occurs when refocusing sparsely sampled light fields.

II. Related Work

In this section we discuss the topics related to our work. We first give a brief review of light field editing and then present existing over-segmentation methods for video, multiview stereo and RGB-D data.

Light field editing: So far, most of light field research has focused on typical light field applications: refocusing [22], depth estimation [23], [24], super-resolution [25], [26]. Light field editing has only been recently addressed with methods targeting either automatic propagation of user inputs from one view to the others [11], or object segmentation.

In the first category, the authors in [11] describe an approach using a 3D voxel-based model of the scene with an associated radiance function to propagate pixel edits and illumination changes in a consistent manner from one view to the other views of the light field. The 3D voxels are then carved away to enforce consistency with an a-priory specified scene radiance model. The voxels whose image projections do not conform well to the local scene radiance model are carved away. Thus, propagating a pixel edit requires determining the voxels that correspond to that pixel and modifying their radiance functions, the change is then propagated by projecting the voxel into each image. The authors in [12] extend the stroke-based 2D image edit propagation method of [27] to light fields. To overcome the computational burden inherent to light field data, the edits are propagated in a downsampled version of the light field. The downsampling step can be seen as some pixel clustering based on an affinity metric defined in the 4D light field space. The pixels are first projected into the affinity space which is then subdivided into clusters, in a way similar to the bounding volume hierarchy. Once the edits have been propagated in the low resolution version of the light field, an upsampling is performed guided by the full resolution data as in the Joint Bilateral Upsampling technique [28]. In [7], the authors present a study on two ways users can interact with light fields. Experiments have been carried out where subjects are asked to perform different typical light field editing tasks, using a ground truth depth map. It is shown that the same tasks can be performed without having a perfect depth map. Using an estimated depth map they observe only a few differences on the capacity of the users to perform the editing. Assuming that light field data can be well approximated by a fixed number of scene layers at different depth, a depth-layer-aware image synthesis method is proposed in [14] for edit propagation.

Perhaps more linked to over-segmentation, the second class of approaches aims at providing object segmentation masks on all views. This can be done using level sets [29]–[31], but this assumes each segmented object to be fronto-planar to the camera. An alternative is to segment each ray using the spatial and angular neighbourhood. In [8], [9], a random forest and SVM technique are used respectively to learn a color and depth model using user scribbles on the central view. A regularisation is then run to obtain the segmentation masks using respectively depth aware and raw angular neighbourhood. To decrease the computational load of the regularisation, it is possible to merge several rays coming from the same scene point in the regularisation [10], but this relies on a relatively accurate dense depth estimation and still scales badly with the spatial light field resolution.

The common point of all these approaches is that they all rely, at some stage, on pixel-level representations of light fields, yielding a high computational load, especially when global regularisation is involved, which motivated the proposed concept of super-rays and can be seen as an extension of super-pixels to light fields.

Super-pixel algorithms: Superpixels have been introduced to circumvent the computational complexity issue in traditional image processing. The term *super-pixel*, first coined in [20] is often described as the partitioning (or clustering) of image pixels into a set of perceptually uniform regions. Ideally super-pixels should be compact (uniform in size), adhere well to the boundaries of objects and be fast to compute.

Because of these properties, super-pixels efficiently represent the image content and are often used as an alternative to pixel representations. Super-pixels allow reducing the computational complexity of many image processing tasks such as object segmentation or object tracking, while providing useful region-based information (e.g texture description or guided regularisation).

Many super-pixel approaches have been proposed and they can be classified into two main categories (see [32] for a recent overview). The first type of methods concerns graph based approaches [33]–[37]. While these methods offer a good accuracy, they either do not provide control on the shape of super-pixels, are very computationally expensive or are not parallelisable, hence not suitable for our applications. The second category of approaches, usually faster than graph-based solutions, aims at growing or evolving existing super-pixels.

This category includes a variety of methods such as the multi-scale watershed segmentation approach proposed in [38], the turbopixels segmenting the image into a lattice-like structure of compact regions by dilating seeds [39], and the quick shift clustering technique [40]. In the latter category, one also finds the SLIC [21] and the SEEDS [32] methods for super-pixels on which we focus in the sequel.

Simple Linear Iterative Clustering (SLIC) super-pixels [21] rely on a reformulation of Loyd's algorithm for the k-means problem with two novelties. First, the distance metric is a weighted sum of the CIELab color distance and the Euclidean pixel distance. Second a search window around the centroid is used to reduce the complexity of the algorithm. This approach has been extended in [41] to take into account a geodesic distance between pixels. Color and spatial information are integrated along the shortest path between two pixels in order to guarantee compactness, color consistency and connectivity.

SEEDS super-pixels [32] take quite the opposite approach. The method starts from a regular, coarse grid segmentation, and iteratively updates blocks of pixels at the edge for the current segmentation. That update is done such that each block can change its super-pixel labelling if it decreases a total energy function of the color distribution of the super-pixels. The block size is reduced along the iterations at a given rate.

Despite the fact that the two last approaches provide better results, as they enforce continuity between super-pixels, in SLIC, the computation for the assignment and update steps can be done for each pixel independently. This is not the case for all approaches relying on any *stack formulation*, which makes the methods awkward to implement in parallel. To the best of our knowledge, only SLIC has been successfully implemented on GPUs [42] to provide results in real time.

Multiview segmentation: Since a light field can be seen as a (possibly dense) collection of views, the proposed work raises issues one can also find when dealing with multiview segmentation. One issue is in particular the possibility of simultaneously perform segmentation and depth estimation.

Super-rays imply establishing correspondences between rays corresponding to different views during the segmentation. The correspondences are found with the help of sparse depth information. While it is possible to simultaneously estimate depth and perform the segmentation [43] or the over-segmentation [44], this can understandably be achieved at the cost of a very high computational cost.

An alternative approach is to first compute super-pixels independently for each view, and then find correspondences between them. In [45], a graph is constructed connecting relatively small super-pixels computed on each view separately. The super-pixels form the vertexes of the graph, while the edges connect super-pixels in all the neighbouring images that satisfy the epipolar constraint. To enforce spatial consistency, the weights on edges are given by the color-consistency of two connected super-pixels. Using a foreground and background color model, the authors are able to extract a foreground object directly. The goal being object segmentation, the super-pixels are not explicitly grouped in a consistent manner across views, as targeted here with the proposed super-rays. In [46], super-pixels are computed on one view only, and are assigned a normal and depth measure using photo-consistency with the other views. The authors are more interested in estimating depth information in a reference view rather than by the correspondence established by view segments.

Video over-segmentation: Over-segmentation has also been studied for reducing the complexity of video analysis tasks. Two main categories of approaches exist for video over-segmentation, either considering a set of consecutive frames as a volume or processing each image separately and updating super-pixels as a new frame arrives.

In [47], super-voxels are computed for a set of consecutive frames. The authors assess five super-voxel algorithms in terms of spatio-temporal coherence, object and region boundary detection, region compression and parsimony. Redundancy between frames in the temporal dimension is hence exploited to construct the super-voxels as it is in the inter-view dimension for multiview data. The authors in [48]–[52] instead try to compute temporally consistent super-pixels. The approaches hence consist in updating the super-pixels as each frame arrives, either by deleting, creating or updating super-pixels to account for the scene motion. Equally, optical flow is often used as an additional clustering information, but also when large displacements are involved.

Although applicable to densely sampled light fields, the first kind of approaches is likely to fail in the case of sparsely sampled light fields as they usually fail for videos in the case of large object displacements. The second type of approaches applies to light fields but does not exploit the fact that object displacements from one view to the other is, due to the scene geometry, uniform.

RGB-D clustering: To be complete, one should also mention the work focusing on RGB-D over-segmentation. However, this problem differs from ours in the sense that the goal is to segment a point cloud rather than pixels on several views. Nevertheless, one paper [53] interestingly uses a modified version of SLIC, using seeds defined from the 3D map, and performs the assignment step using the image distance from the centroid projection in order to circumvent the errors in depth estimation. Note that this work differs from ours as we do not assume dense depth information to be available and, in addition, we target view segmentation rather than point cloud segmentation.

III. Super-ray Light Field Over-Segmentation

Let r be a light ray of the light field LF, and (s,t,x,y) its coordinates using the two plane parametrisation [54], where (s,t) and (x,y) are the angular (view) and spatial (pixel) coordinates respectively. Besides, each light ray has an associated $CIELab$ color value Lab_r. In this work, we note $(x',y') := \mathcal{P}^d_{s',t'}(x,y) \in \mathbb{R}^2$ the spatial pixel position in view (s',t') imaging the same scene point, at a distance d, as (x,y) in view (s,t). This is, (x,y) and $\mathcal{P}^d_{s',t'}(x,y)$ are corresponding points imaging the same scene point in different views. In particular, in the case of a uniformly sampled light field we have

$$\mathcal{P}^d_{s',t'}(x,y) = \big(d\,(s-s') + x,\ d\,(t-t') + y\big). \qquad (1)$$



Fig. 1. Example of angular patches in Ω for a light field of 3×3 views. The orange color corresponds to the reference view (s_c, t_c) so the angular patches are equal to 1 at this position. White positions corresponds to visible rays, so its value is equal to 1, and grey positions are equal to 0. The leftmost patch assumes the ray is visible in all views. Other patches correspond to partial visibility.

However, if the light field has been acquired with a camera array, $\mathcal{P}$ should take into account the extrinsic and intrinsic matrices of each camera, and allow us to estimate the pixel correspondences in this particular setting. Using this notation, $r \sim r'$ are corresponding rays imaging the same scene point, where $r' := (s', t', \mathcal{P}^d_{s',t'}(x, y))$.

Now, given a light field, our goal is to group in the so-called *super-rays*, all perceptually similar rays corresponding to the same scene area. Formally, we aim to compute the mapping $A : LF \subset \mathbb{Z}^4 \rightarrow \mathbb{Z}$, such that each light ray r of the light field is assigned with a super-ray label c. We define SR_c the set of rays r such that $A(r) = c$. Each super-ray SR_c is characterised by a centroid ray r_c. By definition, the angular coordinates of r_c correspond to the fixed *reference view* (s_c, t_c). Besides, each centroid ray has a depth d_c associated to it.

A. Method Description

Initialisation and depth estimation for centroids: First of all, the spatial positions (x_c, y_c) of the centroid rays are initialized on a regular grid of step S in the reference view. The corresponding $CIELab$ color values on such positions are the initial color values of the centroid rays Lab_{r_c}.

Then, a depth d_c is estimated for each centroid ray r_c. As this step is important for the rest of the algorithm the depth estimation needs to be robust. Thus, inspired by the recent works on light field depth estimation [19], we consider a multi-baseline block-matching strategy with angular patches in order to be more robust to occlusions and fattening errors. Let Ω be the set of angular patches where each patch $o \in \Omega$ is defined such that $o(s, t)$ is 1 if a ray is visible on the view (s, t), and 0 otherwise. Each angular patch can be seen as a visibility mask. In practice, we define Ω as a predefined set of angular patches, one patch that corresponds to the full view visibility and eight patches corresponding to different half view visibilities (top-bottom, right-left and diagonals). See an example for a 3×3 light field in Fig. 1. Hence, the depth for the centroid c is estimated by minimizing the color distance in the RGB color space using the different angular patches

$$d_c = \arg\min_d \left\{ \min_{o \in \Omega} \sum_{s',t'} o(s', t')\, \Delta^B_{RGB}(r_c, r'_c) \right\}, \quad (2)$$

where $r'_c = (s', t', \mathcal{P}^d_{s',t'}(x_c, y_c))$ and

$$\Delta^B_{RGB}(r_c, r'_c) = \sum_{(i,j) \in [-B,B]^2} \left(RGB_{r_c}(i, j) - RGB_{r'_c}(i, j) \right)^2, \quad (3)$$

is the patch color distance between the patch in the reference view (s_c, t_c) and the patch in $(s', t') \neq (s_c, t_c)$.

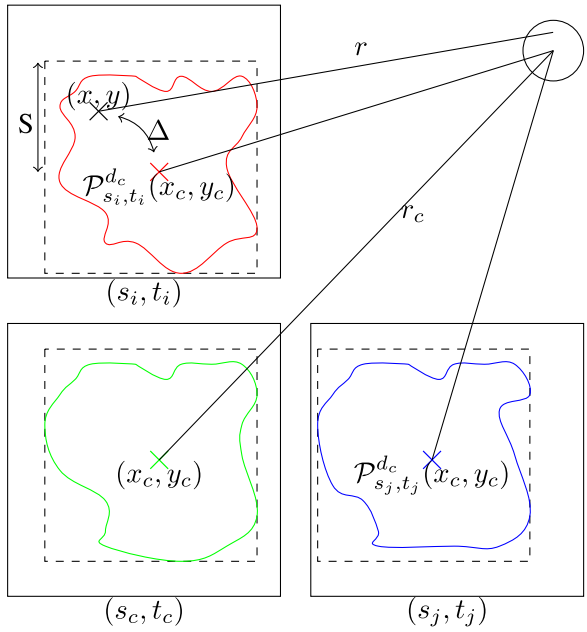


Fig. 2. Assignment step. r is a ray inside the search window of the super-ray SR_c, defined according to the projection of its centroid r_c, $P^{d_c}_{s',t'}(x_c, y_c)$ in the view where r lies. The color and spatial distances in Eq. (4) is denoted Δ.

In particular, $RGB_{r_c}(i, j)$ is the RGB-color value of the ray $(s_c, t_c, x_c + i, y_c + j)$. In this work, we fix $B = 3$ and we consider 9 angular patches (their size being equal to the number of views in the light field). Since the depth is estimated for a few points (the centroids), this choice is acceptable for low complexity applications.

Assignment step: At each iteration, each light ray $r(s, t, x, y)$ of the light field is assigned a super-ray label. First, the depth estimation in the previous step is used to compute the corresponding rays of r_c. Formally, we compute $r'_c = (s', t', \mathcal{P}^{d_c}_{s',t'}(x_c, y_c))$ such that $r_c \sim r'_c$. Then, each ray in a neighbourhood $N_S(r'_c)$ of size S around r'_c, is assigned to the super-ray SR_c if it minimizes the color and spatial distances:

$$A(r) = \arg\min_c \left(\Delta_{Lab}(r, r_c) + m\, \Delta_{xy}(r, r'_c) \right), \quad (4)$$

where

$$\Delta_{Lab}(r, r_c) = ||Lab_r - Lab_{r_c}||^2, \quad (5)$$

$$\Delta_{xy}(r, r'_c) = ||(x, y) - \mathcal{P}^{d_c}_{s',t'}(x_c, y_c)||^2; \quad (6)$$

and m is the parameter weighting the color and spatial distances. A visual explanation can be found in Fig. 2. Note that, when r belongs to the reference view, $r_c = r'_c$ in Eq. (4) and our assignment step is equivalent to the SLIC assignment step. However, our approach allows to coherently assign a label to all rays in the other light field views.

In our assignment step we assume (as in [10]) that two light rays from the same view and close spatial coordinates are likely to image two close scene points. Therefore, a ray that is similar in appearance and close to a centroid light ray or close to one of its corresponding rays is considered likely to belong to the

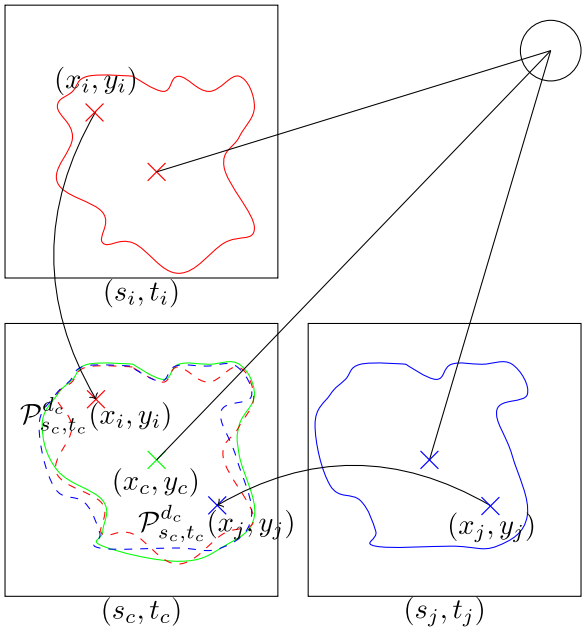


Fig. 3. Update step. Each ray of the super-ray SR_c is reprojected on the reference view using the depth of the super-ray. Here we show r_i and r_j being reprojected on the reference view (s_c, t_c). The projections are averaged, giving the new centroid ray position on the reference view (x_c, y_c).

same scene object. Therefore, it should belong to the super-ray corresponding to this centroid.

Update step: In this step, the spatial coordinates of the ray centroid and its corresponding Lab values are updated. In particular, the new color value of r_c is the average of the color values of all rays in SR_c and the new spatial coordinates are the average coordinates of all light rays, $r = (s, t, x, y)$ in SR_c projected on the reference view using the depth d_c:

$$Lab_{r_c} = \frac{1}{|SR_c|} \sum_{r \in SR_c} Lab_r, \tag{7}$$

$$(x_c, y_c) = \frac{1}{|SR_c|} \sum_{r \in SR_c} \mathcal{P}^{d_c}_{s_c,t_c}(x, y). \tag{8}$$

Note that the centroid rays are defined on a reference view so its angular coordinates (s_c, t_c) are not changed in our algorithm. On the contrary, the centroid spatial coordinates (x_c, y_c) are first initialized on a regular grid in $\mathbb{Z}^2$ and then updated in Eq. (8), which produces new coordinate values in $\mathbb{R}^2$. So, r_c is defined as a virtual light ray which is not necessarily one of the light rays captured in the light field. We summarize the update step in Fig. 3.

When updating the spatial coordinates we assume that rays inside the same super-rays are likely to have similar depth, so Eq. (8) is a good approximation with respect to the centroid position we would obtain using the true depth per ray.

Furthermore, Eq. (8) ensures that two corresponding rays, on two different views, have nearly the same spatial distance Δ_{xy} (as in Eq. (4)) from a given centroid ray. This is not necessarily the case when seeding the centroids independently on all the views.

Algorithm 1: Our super-ray algorithm.

Data: Input Light Field LF
Result: Super-ray assignments A
Initialize centroids on reference view;
while *not (converged OR max. iteration reached)* **do**
  \\ Assignment step
  for *each centroid* c **do**
    for *each view* (s', t') **do**
      Compute $(x'_c, y'_c) = \mathcal{P}^{d_c}_{s',t'}(x_c, y_c)$;
      for *each ray* r *in* $\mathcal{N}_S(r'_c)$ **do**
        $A(r) = c$ (c minimizing Eq. 4);
  \\ Update step
  for *each centroid* c **do**
    Compute $\mathcal{P}^{d_c}_{s_c,t_c}(x, y), \ \forall r \in SR_c$;
    Update Lab_c and (x_c, y_c) (Eqs. 7 & 8);
Cleanup step

Cleanup step: Similarly to SLIC, our algorithm does not enforce super-ray spatial connectivity, so after our light ray grouping procedure some rays may remain isolated, specially when the spatial term in Eq. (4) has a low weight. For this reason, a simple post-processing is performed, that consists in re-labeling super-ray disconnected components (with a number of pixels $< \frac{1}{4}S^2$) with the closest super-ray label.

The entire algorithm proposed in this paper is described in Algorithm 1.

B. Experiments

In order to quantitatively evaluate the proposed approach, well-known super-pixel quality measures can be trivially extended considering all views, such as the Achievable Segmentation Accuracy (ASA), the Boundary Recall (BR) [55] or the Corrected Under-segmentation Error (CUE) [32]. However, these measures do not evaluate the coherence through the light field views. For this reason, we introduce a new evaluation measure called View Consistency (VC). This new measure assumes that the ground truth depth D is known and uses it to select the light rays to consider. Indeed, given a light ray r, our measure aims at evaluating the assignment consistency for the set of corresponding rays imaging the same scene point as r:

$$L'(r) = \Big\{ r' \in LF \ \text{ s.t. } \ (x', y') = \mathcal{P}^{D(r)}_{s',t'}(x, y), \\ (x, y) = \mathcal{P}^{D(r')}_{s,t}(x', y'), \ \ (s', t') \neq (s, t) \Big\}. \tag{9}$$

Note that the re-projection check using the ground truth D takes into account the occlusions and guarantees that $L'(r)$ contains light rays imaging the same scene point. Therefore, we define

$$VC(A) = \frac{1}{|LF|} \sum_{r \in LF} \frac{1}{|L'(r)|} \sum_{r' \in L'(r)} \delta\big(A(r), A(r')\big), \tag{10}$$

where δ is the Kronecker delta. This metric is somehow related to the *Inter-Frame Label Consistency* [49] for super-pixel evaluation in the case of videos, but instead of computing

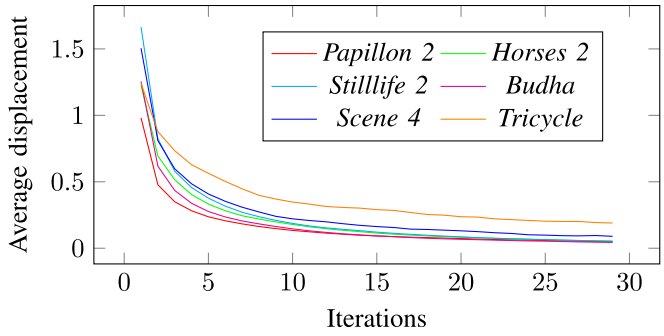


Fig. 4. Average displacement in pixels of the centroid spatial coordinates with respect to the number of iterations.

the consistency from frame to frame using the ground truth optical-flow, we measure the consistency between all light field views simultaneously using the ground truth depth.

Our quantitative evaluation is performed on synthetic datasets, with segmentation and depth ground truth. We use the dataset in [24], that is composed of 9×9 densely sampled views of 768×768 pixels (*Papillon 2, Horses 2, Stilllife 2, Budha*). We also use the dataset in [10], which is a 4×4 sparsely sampled light field views of 640×360 pixels (*Scene 4*). Finally, we propose a new dataset of 5×5 views of 640×480 pixels (*Tricycle*), which is rather sparse.

First, we observe that our approach converges in 10-15 iterations, similarly to SLIC, as shown in Fig. 4, for both dense and sparse light fields.

We compare the proposed super-rays construction method with what we would obtain by separately computing super-pixels on each view, and then merging super-pixels having the highest number of corresponding rays across views. In particular, we use the ground truth depth in the synthetic datasets to re-project rays onto the central view, and we then merge the super-pixels of different views with the super-pixels on the central view having the highest number of re-projected rays, i.e such that VC is maximised. Figs. 5 and 6 show the superiority of our strategy compared to the merging of independent super-rays.

Fig. 5 shows the behaviour of the four quality metrics (ASA, BR, CUE, VC), when varying the different parameters, i.e., the size S of the super-rays and the compactness parameter m, for the dataset *Scene 4*. We observe in Fig. 5 a that when increasing the values of S and m, the super-rays do not segment correctly the objects in the scene, as it was observed with SLIC super-pixels. We also remark that ASA, BR and CUE have similar behaviours with similar numerical values when varying S and m, but the proposed metric VC has an opposite behaviour. Indeed, decreasing S and m decreases the view consistency. This can be explained by the fact that decreasing S and m increases the number of super-rays, hence of super-ray edges near which rays are more prone to labelling errors. So, the view consistency decreases. Fig. 5 b shows that merging independent super-pixels has no impact on the super-pixel metrics (ASA, BR, CUE) as one may expect, but view consistency is severely deteriorated.

The same observations generalises to the rest of our test set. Fig. 6 shows how the two approaches compare when fixing one of the parameters. To be able to compare light fields of different spatial resolutions, we use k, the number of visible super-rays per view. Once again, we observe very close results when changing k and the spatial weighting parameter m for the ASA, BR and CUE. However, enforcing the super-rays to have the same centroid ray imaging the same scene point, yields super-pixel consistency across views. Our approach allows the segmentation consistency to be independent of the parameters k and m, whereas when computing super-pixels on each view, one super-pixel on a view can be described at two (or more) disjoint pieces of super-pixels on another view, depending mostly on the initial seeding. The other thing we notice is the significant difference in terms of over-segmentation performance between densely sampled and sparsely sampled light fields. The over-segmentation of sparsely sampled light fields is less consistent across views, and usually slightly less accurate than for dense light fields. This can be explained by errors in the initial depth estimation, leading to some inconsistent super-rays.

Figs. 7 and 8 show the super-rays constructed by the proposed algorithm, with $m = 1$ for the smoothness parameter and $S = 15$ and $S = 20$ for the super-pixels size respectively. Note that we only display 3×3 views for the sake of readability. Each super-ray is reassigned a random color, the projection of each centroid ray on the different views is represented with a small cross.

At a first glance, the super-rays on each view look like regular super-pixels, but the main difference is that super-rays are overall consistent from a view to another, despite occlusions. We invite the reader to zoom in to see the details.

Regarding the running time, we currently have two implementations, one on CPU using C++ and the other one on GPU using Python and Opencl. None of these implementation is optimised (the GPU implementation uses global memory and atomic operations) but still give low run-time on our laptop equipped with an $Intel\ i7-5600U$ and a $Radeon\ R7-M260X$. On the *Tsukuba* dataset, the super-pixels are computed in 6s and 0.3 s on GPU. For the dataset of [24], we have a run-time of 80 s on CPU and 4.2 s on GPU.[6] For Lytro Illum light field, it takes 57 s on CPU and 3 s on GPU.

We invite the reader to consult our website[7] for more detailed results.

C. Discussion

The main interest of super-rays is to give a compact multi-view representation of the scene, without relying on dense depth estimation. This representation can be used directly or with depth information associated to each centroid ray, as we show in the next section. One main challenge of super-rays is to be as computationally efficient as super-pixels. While it would be possible to integrate some photo-consistency prior in the assignment step (Eq. 4), this would come with a huge computational cost, since the photo-consistency would either need to be pre-computed for each ray and for each candidate disparity value, or

[6] we used 7×7 views on GPU because the entire light field overflowed the maximum allowed buffer size in our implementation.

[7] http://www.irisa.fr/temics/demos/Superrays/index.html

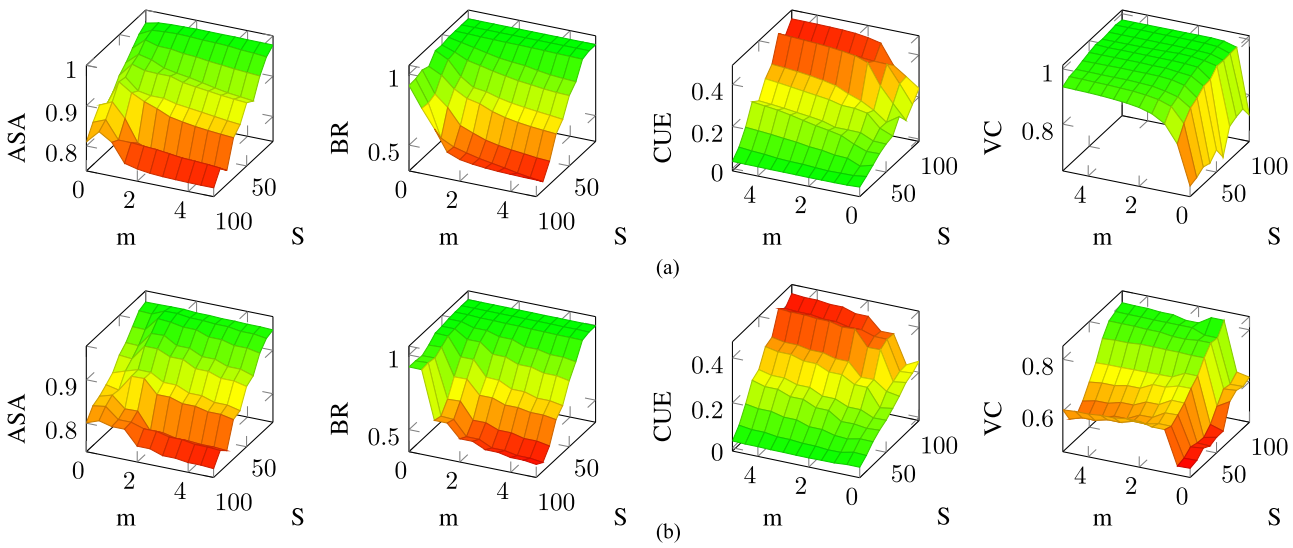


Fig. 5. Different evaluation metrics across different parameters for (*Scene 4*). Green means good score while red represent bad score. For the sake of readability, the axes are flipped differently for each metric. (a) Super-rays evaluation (b) Evaluation of independent super-pixel estimation and *a posteriori* merging.

to be computed at each iteration. Moreover, the K-means strategy applied here relies on some strong assumptions on the data (e.g. spherical distribution variance or uniform cluster size), that get easily violated when dealing with other quantities such as color and spatial information. Instead, our approach only uses the disparity of centroid rays, and lets the spatial distance of the reprojected rays do the grouping. In that sense, the geometric information given by the light field is not fully exploited, but on the other hand, as long as two objects have sufficiently different colors, our approach is still sufficient to yield a good segmentation.

The first obvious limitation of this method is that it relies heavily on the centroid depth initialisation. Even if we propose a method to robustify this initial search, errors may have negative consequences on the output segmentation, rays being assigned to the wrong super-ray. Precisely, this is a problem when the disparity error is greater than the super-ray size, as the centroid would potentially fall outside an object during projection. This being said, light-field depth estimation is an active research topic and our depth estimator could be replaced in the future with another sparse and more accurate method.

The second limitation is related to occlusions. Indeed, because the projected centroids coordinates are not placed regularly, but rather warped according to the scene depth, zones with large occlusion have few, or no nearby centroids projections. If an occlusion is bigger than the search window of the super-ray, rays of a view might not be assigned to any super-rays.

IV. Super-Ray Applications

In this section we propose two examples of editing applications that exploit the super-rays presented in this paper. On the one hand, we present a fast light field segmentation algorithm based on super-rays. On the other hand, we present a novel algorithm for correcting angular aliasing for sparsely sampled light fields.

A. Real Time Interactive Segmentation

Light field segmentation has a high computational complexity [8]–[10] and using super-rays is a good strategy for decreasing it, similarly to temporal super-pixels for video segmentation.

We use the same graph structure introduced in [10] in which rays of the light field are packed in ray bundles in order to decrease the number of nodes of the graph since there is a node per ray bundle instead of a node per light ray. However, it requires a dense depth estimation per view to determine the ray bundles. On the contrary, our proposed strategy uses the super-rays and the centroid depth estimates to create the graph. This is, we build a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ where each node in $\mathcal{V}$ corresponds to one super-ray SR_c and we set an edge in $\mathcal{E}$ between two nodes if the two super-rays share a common border in at least one view (adjacent super-rays). Then, we define the energy

$$\varphi = \sum_{c_i} U(c_i) + s \sum_{\substack{c_i \\ c_j \in \mathcal{N}(c_i)}} P(c_i, c_j), \tag{11}$$

where s is the smoothness parameter, $\mathcal{N}$ the function that indicates if two super-rays are adjacent and U and P the unary and pairwise terms respectively. U uses the super-ray average color probability, defined from a Gaussian mixture learnt from the user input and P is defined as a summation of the view conventional pairwise relationship.

In order to minimize Eq. (11) the multi-label Graph-Cuts algorithm in [59], [60] is applied, using an available implementation.[8] Fig. 9 shows the final segmentation for the *Tsukuba* dataset along with the input scribbles in the reference view and the results in [10] for comparison.

The main interest of using super-rays as a base for segmentation is the gain in running time as summarized in Table I. For the Tsukuba data set, the graph cut applied on super-rays only takes

[8]http://vision.csd.uwo.ca/code/

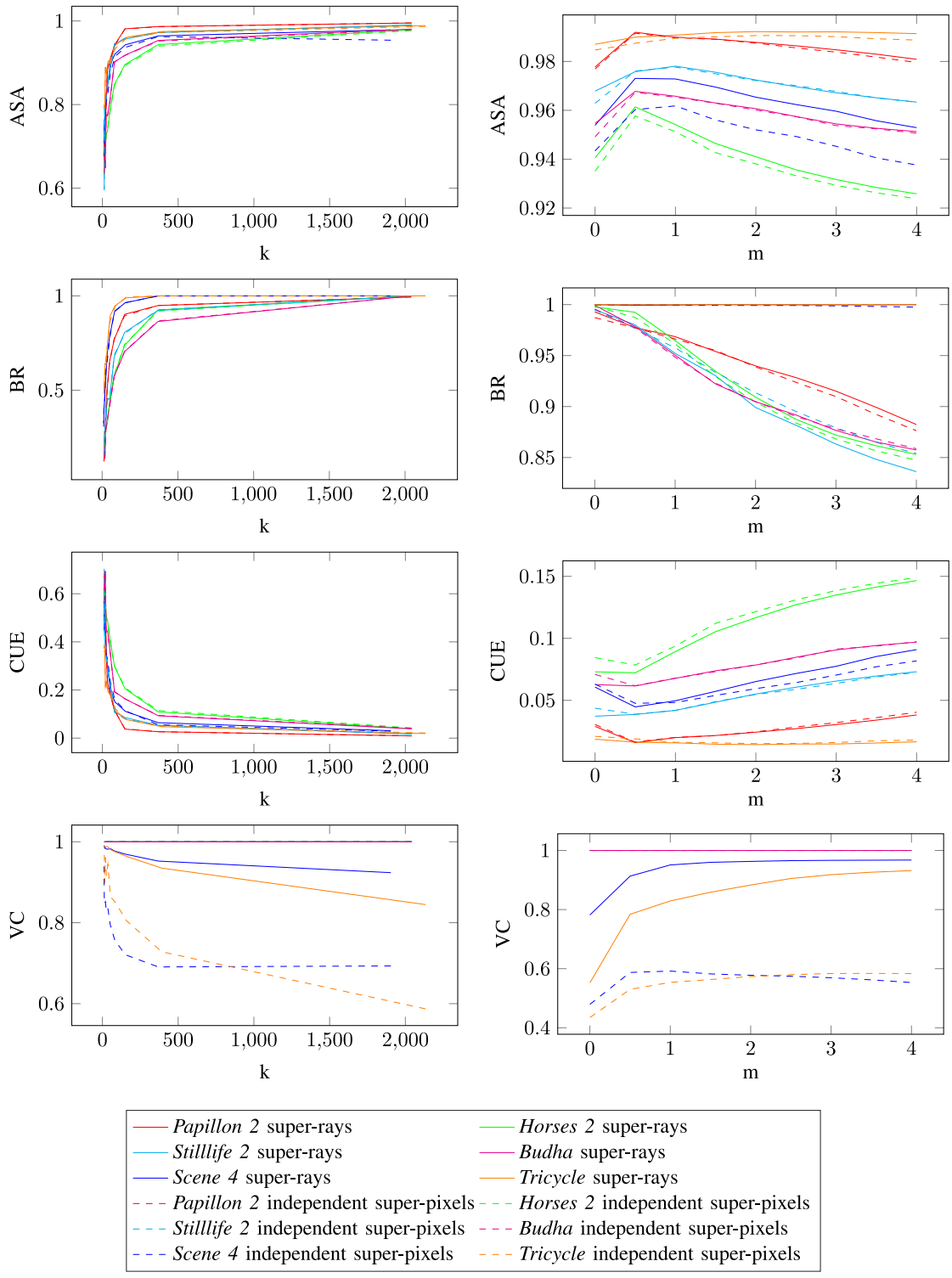


Fig. 6. Comparison of super-rays versus merged independent super-pixels, when varying k, the number of super-rays visible on each view with $m = 1$ fixed (left column) and when varying m the compactness parameter, with $k = 500$ fixed (right column).

1 ms on the same machine as in Section III-B (a laptop with an $Intel\ i7-5600U$ and a $Radeon\ R7M260X$). The complete segmentation algorithm takes 8.1 s on CPU with the following running times for each step: 6 s to compute the super-rays, 2.1 s to build the graph (i.e to compute the models, unary and pair-wise costs), 1 ms to do the graph cut. With super-rays computed on GPU, the segmentation only takes 2.4 s with 0.3 s for the super-rays computation, 2.1 s to build the graph, and 1 ms to do the graph cut. On the dataset in [8] the approach takes 94 s (80 s for the super-rays, 14 s to build the graph, 4 ms for the graph cut) on CPU and 18.2 s (4.2 s for the super-rays, 14 s to build the graph, 4 ms for the graph cut) on GPU.

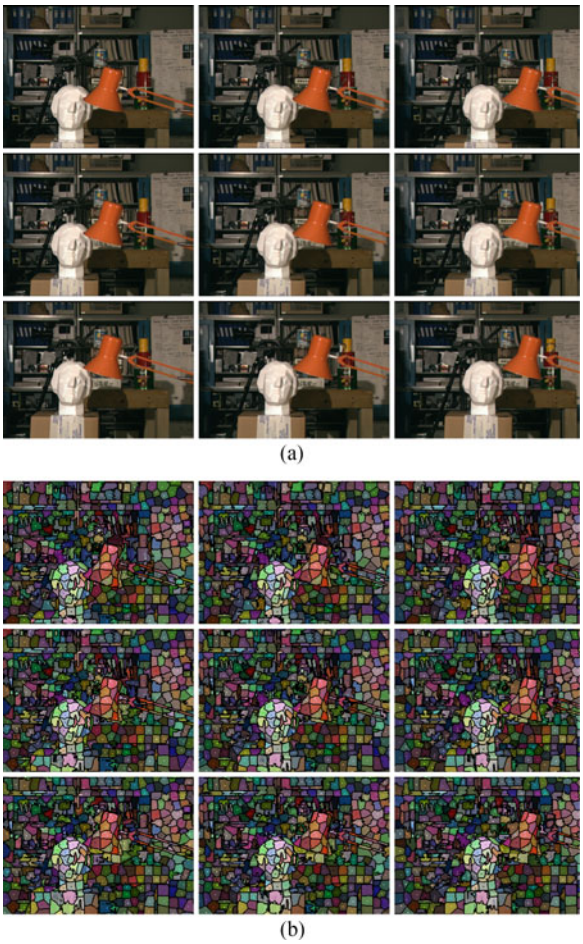


Fig. 7. Super-rays for the sparsely sampled light field in the *Tsukuba* dataset [56]. (a) Original light field (b) Color-coded super-rays.

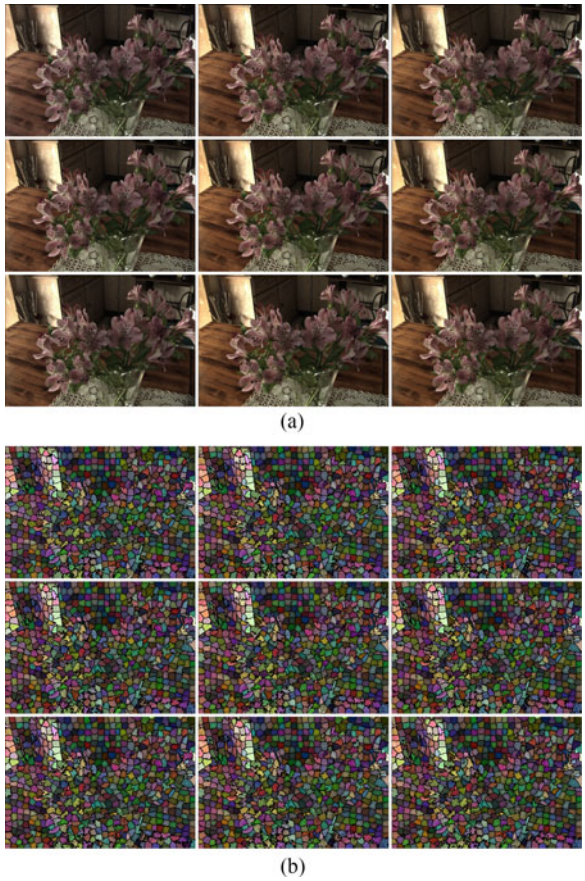


Fig. 8. Super-rays for the densely sampled light field we have captured with the Lytro Illum and decoded using the toolbox in [57], [58]. (a) Original light field. (b) Color-coded super-rays.

These running times are significantly reduced compared with the state-of-the-art methods in [8] and [10]. The authors in [8] report a segmentation time between 6 and 10 min, including 1 to 5 min to learn the models and 5 min to perform the optimization on a more powerful desktop GPU ($Nvidia\ GTX-580$). The approach in [10] which uses depth to reduce the number of nodes in the graph takes 9.3 s on the Tsukuba dataset with the same hardware. It takes 6.3 s to build the graph and 3 s to do the segmentation. On the dataset in [8] and using real depth (in contrast to the ground truth as described in the paper), the approach in [10] takes 122.6 s (82 s to build the graph, 40 s to do the segmentation).

The above figures for the reference methods in [8] and [10] do not take into account the time needed to compute the dense depth map on each view, while our super-ray construction method includes the coarse depth estimation. For a fair comparison, the time needed to estimate the dense depth maps should be added to the above timing reported for [8] and [10] which respectively takes 2.9 s and 33.6 s for the Tsukuba data set and the data set in [8] using the approach in [61] (implemented as pre-processing on a desktop GPU using C++/OpenCL).

In conclusion, in the context of interactive segmentation, we see that avoiding to compute dense depth maps and running the segmentation on a much more compact light field representation allow a significant reduction of the segmentation running time on CPU, and thanks to the parallelizable nature of super-rays, a dramatic reduction when using GPU. We are confident that more implementation efforts will lead the super-rays computation to be near real-time.

This gain comes at the expense of losing precision, but the obtained accuracy is sufficient for many real-time applications requiring a fast (rather than accurate) segmentation. We illustrate this loss of accuracy in Table II. We see that super-rays introduce errors with a rate of 0.1% to 2% in the best case scenario, when the algorithm used to assign a label to each super-ray does not do any error. Using the graph structure and the data and pairwise terms in [10], the accuracy drops, with an error rate of 1% to 3%, mostly due to the limited data cost model. Using a more complex model, using textural information provided by super rays will solve this issue.

B. *Correcting Angular Aliasing*

One of the major new light field applications compared to conventional cameras is the post-capture image refocusing. In fact, light fields captured with Plenoptic type 1.0 cameras provide quite impressive refocusing results [1] but angular aliasing appears when refocusing light fields from camera rigs or refocused

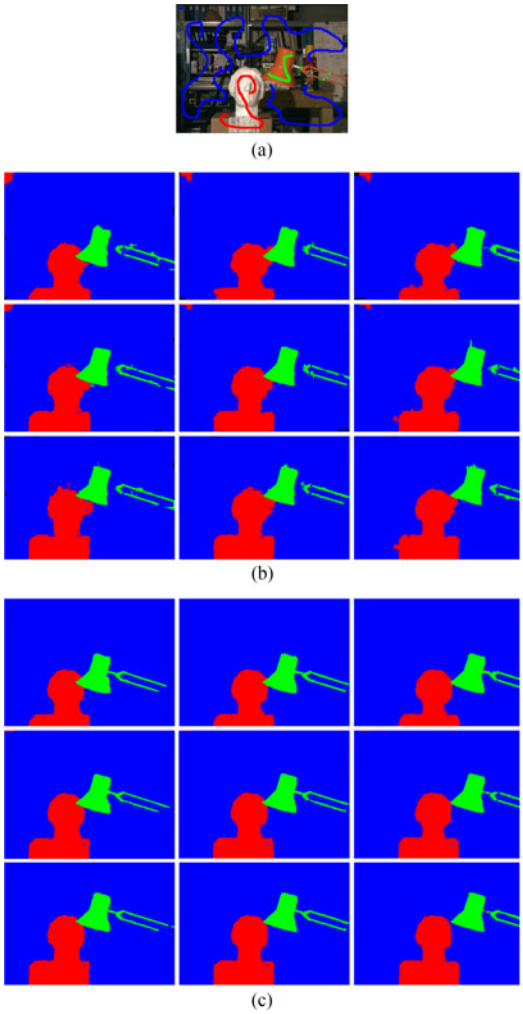


Fig. 9. Graph-cut segmentation of the dataset *Tsukuba* using our super-rays. Our strategy allows a significant reduction of the running time compared to the state of the art light field segmentation. (a) Reference view with input scribbles. (b) Our output segmentation masks using super-rays. (c) Output segmentation masks in [10].

TABLE I
COMPARISON OF THE TOTAL SEGMENTATION RUNNING TIMES, INCLUDING DEPTH ESTIMATION IF ANY

Approach \ Dataset	Tsukuba	[8] dataset
[8] (GPU)	N/S	393.6 to 633.6s
[10] (CPU)	12.2s	156.2s
Ours (CPU)	8.1s	94s
Ours (GPU)	2.4s	14s

Results of [8] were computed on a more powerful machine than the others.

plenoptic cameras. This is due to the poor angular sampling of such acquisition systems. Angular aliasing is particularly visible when the simple *shift-and-add* algorithm is used for refocusing, whereas other solutions such as the adaptive splatting [62] or the rendering of novel views [6], [63] decrease the artifacts.

However, these solutions are depth-based methods which makes them unsuitable for fast editing applications. Here we propose to refocus using our super-rays to avoid angular aliasing while maintaining a low complexity. Our goal is not to render new views but our philosophy is similar in the sense that we use approximate intermediate views computed with our super-rays. Thus, refocusing is performed via *shift-and-add* using both original and approximate intermediate views that we call *virtual views*.

TABLE II
SEGMENTATION ACCURACY (RATIO OF PIXEL ASSIGNED TO THE RIGHT LABELS ACCORDING TO THE GROUND TRUTH) WITH A SYNTHETIC DATASET

Dataset:	Still life 2	Papillon 2	Horses 2	Budha
Result in [8]:	99.3	99.4	99.3	98.6
Result in [10]:	99.2	99.5	99.1	99.1
Our results (GT Segmentation)	99.0	99.5	97.0	98.1
Our results (Graph cut)	98.8	99.1	96.1	96.6

The parameters were fixed to $k = 2000$ and $m = 1$. We first show our results with respect to the dominant ground truth label assignment for the segmentation (as when computing the ASA metric). This allows us to measure the loss in segmentation accuracy solely introduced by the super-rays (i.e assuming the segmenting process does not introduce additional errors). We also give the results using the super-rays followed by the graph cut approach of [10].



Fig. 10. Refocusing the sparsely sampled *Tsukuba* dataset. Left: Shift-and-add algorithm which creates annoying angular aliasing artifacts. Right: Our strategy using real and virtual views. The virtual views have been rendered with our super-rays. We invite the reader to zoom-in to see the details.

With our notations, the *shift-and-add* method to compute the refocused image at depth d from the viewpoint (s_c, t_c) is defined as

$$I^d_{s_c,t_c}(x,y) = \frac{1}{N^2}\sum_{s,t} RGB(s,t,\mathcal{P}^d_{s,t}(x,y))\,, \qquad (12)$$

where the original angular coordinates $s, t = 0, \ldots, N-1$; correspond to the original image views available in the light field. Now, we define the virtual positions $(u, v) \in \mathbb{R}^2$ as

$$(u,v) = (s,t) + \frac{1}{\Delta_d}(m,n)\,, \quad m,n = 1,\ldots,\Delta_d - 1\,, \quad (13)$$

where the number of virtual views between two original views is $\Delta_d = [|(d_c - d)|]$ the integer part of the absolute value of the depth difference between the refocusing depth d and the centroid depth d_c, c being the centroid label to which (s_c, t_c, x, y) belongs. Therefore, the corrected refocused image is defined as the sum of the original images and the virtual views contribution:

$$I'^d_{s_c,t_c}(x,y) = \frac{1}{(N\cdot\Delta_d)^2}\sum_{s,t}\Big\{RGB(s,t,\mathcal{P}^d_{s,t}(x,y)) + \sum_{u,v} RGB(u,v,\mathcal{P}^d_{u,v}(x,y))\Big\}. \qquad (14)$$



Fig. 11. Angular aliasing compassion. We show the images refocused with the 2×2 corner views in the test set in [63], the full 8×8 light field (ground truth), the 8×8 synthesized light field with the approach in [63] and the 8×8 light field synthesized with our approach. We use $k = 1000$ and $m = 1$.

Note that Eq. (14) requires to interpolate the virtual views, which can be quite memory and time consuming. However, in our strategy we do not explicitly compute them but we use the super-rays depth d_c to approximate ray colors in virtual views:

$$RGB(u, v, \mathcal{P}^d_{u,v}(x, y)) \simeq RGB([u], [v], \mathcal{P}^{d_c}_{[u],[v]}(\mathcal{P}^d_{u,v}(x, y))), \tag{15}$$

where $([u], [v])$ is the closest original view.

Considering d_c as the depth of all rays of a super-ray is a coarse approximation but it has few or no consequences on the final result since we only approximate for the blurred, out of focus areas. However, the high frequencies on the blur due to the poor angular sampling are successfully removed. Moreover, using the approximate depth information is sufficient to prevent out of focus rays coming from the background to be mixed with occluding in-focus rays which may create artifacts.

Fig. 10 shows the angular aliasing correction using the dataset from [56] when refocusing at the sculpture bust. Our method successfully decreases the angular aliasing on the background. Note that we simulate a squared aperture which provides a squared *bokeh*.

Finally, we compare our approach with the one presented in [63], which presents a Convolutional Neural Network approach to learn view synthesis for light fields. It provides excellent results, with quite big computational cost both for training but also for generating the views (the authors report 12 s to generate a single 541×376 view). Their training and test set consist of Lytro Illum light fields. Only the four corner views are kept, giving a sparse light field with a very small number of views, and the 60 views in between these images are generated to extrapolate an 8×8 light field which is compared to the captured Lytro Illum ground truth. We use our approach on the 2×2 corner views to generate the other views the same way. The parameters used were $k = 1000$ and $m = 1$. The computation of the super-rays takes 0.36 s (0.17 s per iteration) and generating a view takes 0.007 s. Then, we compare on Fig. 11 the refocused image obtained with the *shift-and-add* algorithm using the 2×2 views, the synthesized views from [63] and our synthesised views using the super-rays. We also show the refocused ground truth using the 8×8 Lytro Illum real views.

Quantitatively, on the test set, the average SSIM (Structural Similarity) and PSNR (Peak Signal to Noise Ratio) between the ground truth refocused image and our approach is respectively 0.98 and 39.0 versus 0.99 and 42.6 using [63] and 0.92 and 34.5 for the image refocused using only the 2×2 views. Note that we had to crop a 22 pixels border on the reference and our refocused images to comply with the output of [63] that looses the border pixels. Overall, our results are of lower quality because of depth estimates, but in the zones where depth is correct, we see no or few differences between our result and the ground truth. However, our approach has a significantly lighter complexity compared to [63] (the authors report 10 s to generate a single view using Matlab and CuDNN, while our approach runs in the order of $\frac{1}{100}s$) which is the main interest of our super-rays.

V. Conclusion

We have introduced the concept of super-ray and we have proposed an implementation that is suitable for sparsely and densely sampled light fields. Our solution is fast, easy to implement, and suitable for GPU implementations. Besides, our super-rays are view consistent, which is a major advantage for light field editing. In particular, we have shown the interest of super-rays for light field segmentation and refocusing without

angular aliasing. In the future, the proposed approach could be used for other light field editing tasks, such as intrinsic decomposition [64] or video light field processing.

REFERENCES

[1] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Comput. Sci. Tech. Rep.*, vol. 2, no. 11, pp. 1–11, 2005.

[2] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *Proc. IEEE Int. Conf. Comput. Photogr.*, 2009, pp. 1–8.

[3] C. Zhang and T. Chen, "A self-reconfigurable camera array," in *Proc. SIGGRAPH Sketches*, 2004, p. 151.

[4] B. Wilburn, N. Joshi, V. Vaish, M. Levoy, and M. Horowitz, "High-speed videography using a dense camera array," in *Proc. IEEE Comput. Soc. Comput. Vision Pattern Recog.*, vol. 2, 2004, pp. II-294–II-301.

[5] K. Venkataraman *et al.*, "Picam: An ultra-thin high performance monolithic camera array," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 166-1–166-13, 2013.

[6] C.-T. Huang, J. Chin, H.-H. Chen, Y.-W. Wang, and L.-G. Chen, "Fast realistic refocusing for sparse light fields," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 1176–1180.

[7] A. Jarabo, B. Masia, A. Bousseau, F. Pellacini, and D. Gutierrez, "How do people edit light fields?" *ACM Trans. Graph.*, vol. 33, no. 4, 2014. [Online]. Available: http://www-sop.inria.fr/reves/Basilic/2014/JMBPG14

[8] S. Wanner, C. Straehle, and B. Goldluecke, "Globally consistent multi-label assignment on the ray space of 4D light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 1011–1018.

[9] H. Mihara, T. Funatomi, K. Tanaka, H. Kubo, H. Nagahara, and Y. Mukaigawa, "4D light-field segmentation with spatial and angular consistencies," in *Proc. IEEE Int. Conf. Comput. Photogr.*, 2016, pp. 1–8.

[10] M. Hog, N. Sabater, and C. Guillemot, "Light field segmentation using a ray-based graph structure," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 35–50.

[11] S. M. Seitz and K. N. Kutulakos, "Plenoptic image editing," *Int. J. Comput. Vis.*, vol. 48, no. 2, pp. 115–129, 2002.

[12] A. Jarabo, B. Masia, and D. Gutierrez, "Efficient propagation of light field edits," *Proc. SIACG*, 2011, pp. 75–80.

[13] H. Ao *et al.*, "Light field editing based on reparameterization," in *Proc. Pac. Rim Conf. Multimedia*, 2015, pp. 601–610.

[14] F.-L. Zhang, J. Wang, E. Shechtman, Z.-Y. Zhou, J.-X. Shi, and S.-M. Hu, "Plenopatch: Patch-based plenoptic image manipulation," *Trans. Vis. Comput. Graph.*, 2016, pp. 1561–1573.

[15] K. W. Shon *et al.*, "Spatio-angular consistent editing framework for 4D light field images," *Multimedia Tools Appl.*, vol. 75, pp. 1–17, 2016.

[16] C. Chen, H. Lin, Z. Yu, S. Bing Kang, and J. Yu, "Light field stereo matching using bilateral statistics of surface cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1518–1525.

[17] H. Lin, C. Chen, S. Bing Kang, and J. Yu, "Depth recovery from light field using focal stack symmetry," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3451–3459.

[18] W. Williem and I. Kyu Park , "Robust light field depth estimation for noisy scene with occlusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 4396–4404.

[19] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3487–3495.

[20] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 10–17.

[21] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[22] R. Ng, "Fourier slice photography," *ACM Trans. Graph.*, vol. 24, pp. 735–744, 2005.

[23] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 673–680.

[24] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4D light fields," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2012, pp. 41–48.

[25] T. E. Bishop, S. Zanetti, and P. Favaro, "Light field superresolution," in *Proc. IEEE Int. Conf. Comput. Photogr.*, 2009, pp. 1–9.

[26] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 606–619, Mar. 2014.

[27] X. An and F. Pellacini, "Appprop: All-pairs appearance-space edit propagation," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 40-1–40-9, 2008.

[28] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, 2007.

[29] J. Berent and P. L. Dragotti, "Unsupervised extraction of coherent regions for image based rendering," in *Proc. Brit. Mach. Vis. Conf.*, 2007, pp. 1–10.

[30] P. L. Dragotti and M. Brookes, "Efficient segmentation and representation of multi-view images," in *Proc. SEAS-DTC Workshop*, Edinburgh, Scotland, 2007.

[31] J. Berent and P. L. Dragotti, "Plenoptic manifolds—Exploiting structure and coherence in multiview images," *Signal Process. Mag.*, vol. 24, pp. 34–44, 2007.

[32] M. Van den Bergh, X. Boix, G. Roig, and L. Van Gool , "Seeds: Superpixels extracted via energy-driven sampling," *Int. J. Comput. Vis.*, vol. 111, no. 3, pp. 298–314, 2015.

[33] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[34] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[35] A. P. Moore, S. J. Prince, J. Warrell, U. Mohammed, and G. Jones, "Superpixel lattices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[36] O. Veksler, Y. Boykov, and P. Mehrani, "Superpixels and supervoxels in an energy optimization framework," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 211–224.

[37] Y. Zhang, R. Hartley, J. Mashford, and S. Burn, "Superpixels via pseudo-boolean optimization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1387–1394.

[38] F. Meyer and P. Maragos, "Multiscale morphological segmentations based on watershed, flooding, and eikonal PDE," in *Proc. Int. Conf. Scale-Space Theories Comput. Vis.*, 1999, pp. 351–362.

[39] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "Turbopixels: Fast superpixels using geometric flows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2290–2297, Dec. 2009.

[40] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 705–718.

[41] P. Wang, G. Zeng, R. Gan, J. Wang, and H. Zha, "Structure-sensitive superpixels via geodesic distance," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 1–21, 2013.

[42] R. Birkus, "Accelerated gSLIC for superpixel generation used in object segmentation," *Proc. 19th Central Eur. Semin. Comput. Graph.*, vol. 15, 2015.

[43] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha, "Object stereojoint stereo matching and object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3081–3088.

[44] Y. Taguchi, B. Wilburn, and C. L. Zitnick, "Stereo reconstruction with mixed pixels using adaptive over-segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[45] N. D. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Automatic object segmentation from calibrated images," in *Proc. IEEE Conf. Vis. Media Prod.*, 2011, pp. 126–137.

[46] B. Mičušík and J. Košecká, "Multi-view superpixel stereo in urban environments," *Int. J. Comput. Vis.*, vol. 89, no. 1, pp. 106–119, 2010.

[47] C. Xu and J. J. Corso, "Evaluation of super-voxel methods for early video processing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1202–1209.

[48] A. Levinshtein, C. Sminchisescu, and S. Dickinson, "Spatiotemporal closure," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 369–382.

[49] J. Chang, D. Wei, and J. W. Fisher, "A video representation using temporal superpixels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2051–2058.

[50] M. Reso, J. Jachalsky, B. Rosenhahn, and J. Ostermann, "Temporally consistent superpixels," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 385–392.

[51] M. Van den Bergh, G. Roig, X. Boix, S. Manen, and L. Van Gool , "Online video seeds for temporal window objectness," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 377–384.

[52] M. Reso, J. Jachalsky, B. Rosenhahn, and J. Ostermann, "Fast label propagation for real-time superpixels for video content," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 902–906.

[53] J. Yang, Z. Gan, K. Li, and C. Hou, "Graph-based segmentation for RGB-D data using 3-D geometry enhanced superpixels," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 927–940, May 2015.

[54] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Tech.*, 1996, pp. 43–54.
[55] P. Neubert and P. Protzel, "Superpixel benchmark and comparison," in *Proc. Forum Bildverarbeitung*, 2012, pp. 1–12.
[56] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 7–42, 2002.
[57] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1027–1034.
[58] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Linear volumetric focus for light field cameras," *ACM Trans. Graph.*, vol. 34, no. 2, Feb. 2015, Art. no. 15.
[59] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[60] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.
[61] V. Drazic and N. Sabater, "A precise real-time stereo algorithm," in *Proc. 27th Conf. Image Vis. Comput. New Zealand*, 2012, pp. 138–143. [Online]. Available: http://doi.acm.org/10.1145/2425836.2425867
[62] J. Fiss, B. Curless, and R. Szeliski, "Refocusing plenoptic images using depth-adaptive splatting," in *Proc. IEEE Int. Conf. Comput. Photogr.*, 2014, pp. 1–9.
[63] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 193-1–193-10, 2016.
[64] E. Garces, J. I. Echevarria, W. Zhang, H. Wu, K. Zhou, and D. Gutierrez, "Intrinsic light field images," *Comput. Graph. Forum*, 2017.



Matthieu Hog received the B.Sc. degree in 2013 from the Université de Saint-Etienne, Saint-Etienne, France, and the M.Sc. degree in 2015 from the Université de Saint-Etienne and the Gøjik University College, Norway. He is currently working toward the Ph.D. degree in INRIA (Institut National de Recherche en Informatique et en Automatique) and Technicolor Research & Innovation in Rennes, France. His research interests include image processing, computer vision, and computational imaging.



Neus Sabater received the B.Sc. degree in 2005 from the Universitat de Barcelona, Spain, and the M.Sc. and Ph.D. degrees in 2006 and 2009, respectively, in image processing from the Ecole Normale Supérieure de Cachan, France. She was a Postdoctoral Researcher with the California Institute of Technology before being appointed at Technicolor Research & Innovation in 2011, where she is currently a Senior Scientist. Her research interests include image processing, computer vision, and computational photography.



Christine Guillemot (F'13) received the Ph.D. degree from Ecole Nationale Supérieure des Telecommunications (ENST), Paris, in 1992. She is currently the Director of Research at Institut National de Recherche en Informatique et Automatique (INRIA), France. From 1985 to 1997, she has been with France Télécom in the areas of image and video compression for multimedia and digital television. From 1990 to mid-1991, she has worked as a Visiting Scientist at Bellcore Bell Communication research, USA. Her research interests are signal and image processing, and in particular 2D and 3D image and video coding, joint source and channel coding for video transmission over the Internet and over wireless networks, and distributed source coding. She has served as Associate Editor for IEEE TRANSACTION ON IMAGE PROCESSING (from 2000 to 2003), for IEEE TRANSACTION ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (from 2004 to 2006), and for IEEE TRANSACTION ON SIGNAL PROCESSING (2007–2009). She is currently an associate editor of the *Eurasip journal on image communication* (since 2010), for the IEEE TRANSACTION ON IMAGE PROCESSING (2014–2016), and for the IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING (since 2013). She has been a member of the IEEE IMDSP (2002–2007) and IEEE MMSP (2005–2008) technical committees. She is currently a member of the IEEE IVMSP-Image Video Multimedia Signal Processing-technical committee (since 2013). She is the coinventor of 24 patents, she has coauthored 9 book chapters, 62 international journal publications, and around 150 articles in peer-reviewed international conferences.

Designing Three-Dimensional Cellular Automata Based Video Authentication With an Optical Integral Imaging Generated Memory-Distributed Watermark

Xiao-Wei Li, Seok-Tae Kim, and Qiong-Hua Wang

***Abstract*—Integral imaging for image watermarking is attracting a great deal of interest as it provides high processing speed and data redundancy property. Traditional video watermarking algorithms are mostly based on discrete transform domains, such as the discrete cosine transform and discrete wavelet transform. Most of these algorithms generally provide one transform plane for watermarking. In this paper, we propose a video watermarking by employing three-dimensional (3-D) cellular automata (CA) transform and perfect-reconstruction integral imaging (PRII). Compared to the existing works, the main strength of this paper is to utilize a 3-D CA transform to produce multiple transform planes for watermark embedding, so that greatly improve the security; decompose watermark into a series of elemental images using integral imaging to enhance the robustness; and employ perfect-reconstruction filter into the iterative back-projection process thus resolving the ringing artifacts problems. Experimental results show that the proposed PRII-based watermarking method outperforms other similar watermarking methods and is robust to a wide range of attacks, e.g., Gaussian noise, salt and pepper noise, frame clipping, JPEG compression, and median filter attack.**

***Index Terms*—3D cellular automata, integral imaging, video watermarking.**

I. Introduction

THREE-DIMENSIONAL (3D) sensing and imaging have been subject of research due to their diverse benefits and applications. Integral imaging is a 3D imaging technique which uses a microlens array or a pinhole array to capture light rays emanating from an object from different directions [1]–[3]. 3D integral imaging technique has attracted much attention in various applications. Integral imaging as a 3D imaging method is considered to be one of the most important technologies for 3D image processing. Integral imaging with combinations of digital image processing is gaining greater attention recently as it can provide high security and computationally simple and feasible. The growth of digital multimedia and Internet technologies has made copyright protection and data encryption become two important issues in the digital world [4]–[12]. Watermarking has been presented to protect the copyright of multimedia by embedding secret information robustly into digital products for many years. A general video watermarking system has two main processes: watermark embedding and watermark extraction. In the embedding process, a watermark usually contains a payload data stream is embedded into an original host video sequence by imposing imperceptible changes to generate the watermarked video signal [13]. Detection involves producing an estimate of the watermark payload from a possibly tampered version of the watermarked video signal. The existing video watermarking techniques are mainly divided into two groups: spatial domain and transform domain in which the watermark data is embedded. The conventional transform-based methods depend on discrete cosine transform (DCT), discrete wavelet transform (DWT), and discrete Fourier transform (DFT) [14]–[23]. Unlike most of these methods, in using cellular automata (CA) [24]–[28] to transform images data, the benefit gained in utilizing a different set of CA gateway values can provide a series of transform planes for watermark embedding.

For video watermarking, another important requirement is to enhance the robustness of a watermark against attacks. Recently, integral imaging is used as a potential tool for image encryption, and watermarking and compression [29]–[33]. Integral imaging divides an input image into many elemental images using an optical lenslet array or a pinhole array, and these elemental images greatly improve robustness and security owing to the memory distribution properties of each elemental image. A general integral imaging system consists of two processes: pickup and reconstruction. In pickup, light rays emanate from images through each lenslet that are captured by an image sensor. These captured demagnified images are regarded as elemental image array (EIA). Reconstruction is generally achieved with optical integral imaging reconstruction (OIIR) and computational integral imaging reconstruction (CIIR).

To improve the OIIR problems of existing limitations such as quality degradation of reconstructed images caused by light diffraction and physical devices, CIIR was developed to increase

Manuscript received December 1, 2016; revised April 3, 2017; accepted May 3, 2017. Date of publication June 12, 2017; date of current version October 23, 2017. This work was supported in part by the National 973 Program of China under Grant 2013CB328802, and in part by the NSFC under Grants 61535007 and 61505127. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Yebin Liu. *(Corresponding author: Qiong-Hua Wang.)*

X.-W. Li and Q.-H. Wang are with the School of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China (e-mail: xwli@scu.edu.cn; qhwang@scu.edu.cn).

S.-T. Kim is with the Department of Information and Communications, Pukyong National University, Busan 608-737, South Korea (e-mail: setakim@pknu.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2714838

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

the resolution of the reconstructed images. CIIR effectively improves the image resolution because of no diffraction and limited physical device. However, it still has some problems, such as the fact that the image resolution seriously degrades as the reconstruction distance increases. Because CIIR is a pixel reverse magnified reconstruction technique, the interference of adjacent pixels degrades the quality of the reconstructed image. Many researchers have focused on improving the quality of the reconstructed image in integral imaging systems. However, many techniques still use the CIIR or the improved version of CIIR, the degradation of the reconstructed image resolution has not effectively resolved.

Recently, to resolve the low-resolution problem, the super-resolution reconstruction technique was employed in the integral imaging system. $A.Stern$ and $B.Javidi$ [34] have presented the time-division multiplexed computational integral imaging, the elemental images are captured by means of time-division multiplexing with a moving lenslet array. In their reconstruction process, the iterative back-projection (IBP) [35], [36] based super-resolution technique was employed to improve the low-resolution of the reconstructed image. A moving least square (MLS) [37] based super-resolution integral imaging method has been presented in [38], in which the MLS-based super-resolution method was first utilized to improve the low-resolution of captured elemental images, and CIIR was performed on these high-resolution elemental images to obtain the final image. This method improves the image-quality. However, the problem of the pixel inverse amplification resulting in pixel overlapping has not been well resolved in CIIR. Furthermore, the high computational cost of MLS-based super-resolution method will aggravate the video watermarking computational burden [39].

The principal strength of 3D CA-based video watermarking is the large numbers of CA transform bases available and the parallel and integer-based character of the computational process involving states of the cellular automata. This can translate into an enormous computational speed in a well-designed while other conventional transform techniques cannot. In this work, we discuss a method for video watermarking by use of the pseudo 3D CA transform and the perfect-reconstruction integral imaging (PRII) technique to cope with these problems. This proposed watermarking method has the following advantages: CA provides multiple transform planes for watermark signal hiding, to the best of our knowledge, 3D CA transform is the first time to be used for video watermarking; CA transform also can significantly reduce the computation burden owing to the inherent parallelism of CA computing; perfect-reconstruction filter employed in the iterative back-projection process to effectively remove the ringing artifacts problems; it outperforms the conventional methods in terms of robustness and imperceptibility with the help of PRII and the fuzzy interference system. The proposed watermarking system is described in Section II where the motivations behind the use of PRII and 3D CA transform profile are outlined. The performance of the watermarking system is evaluated in Section III, where its robustness against benchmark attacks is assessed. Finally, the conclusion is presented in Section IV.

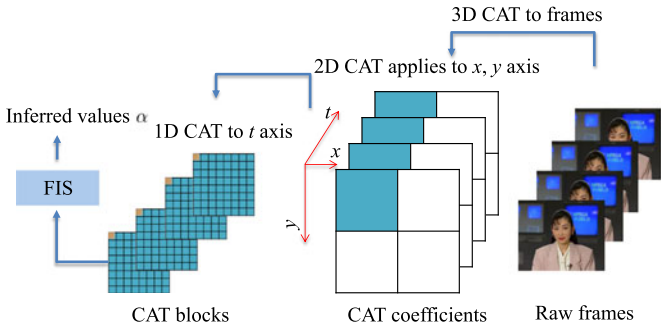


Fig. 1. Illustration of a pseudo 3D CA transform based fuzzy system.

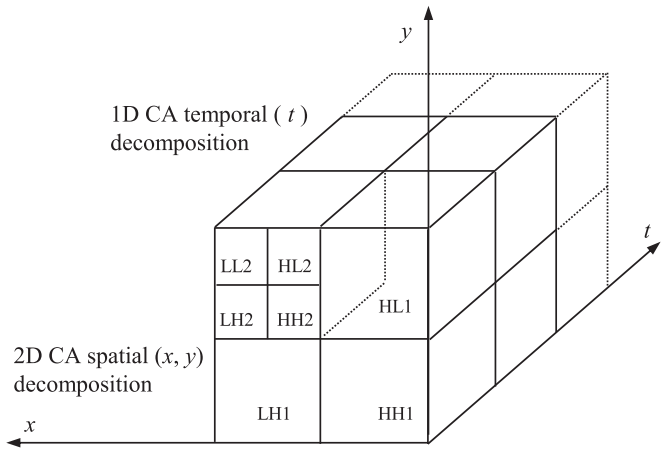


Fig. 2. 3D CA decomposition in spatio-temporal axis.

II. Theoretical Analysis

A. Embedding Domain Selected by 3D CA Transform

To perform the spatio-temporal based watermarking, the video is viewed as a 3D signal with 2D in space and 1D in time. Next, the video frames are decomposed by 3D CA transform filter. Fig. 1 describes the principle of fuzzy-based pseudo 3D CA transform and the detailed explanation of the inferred process will be found in Section II-C. Video data can be treated as 3D signal, each pixel data $I(i,j,t)$ at the point (i,j) and the given time t can be denoted as

$$I(i,j,t) = \sum_{k=0}^{N_k-1} \sum_{l=0}^{N_l-1} \sum_{m=0}^{N_m-1} c_{klm} A_{ijtklm} \tag{1}$$

where A_{ijtklm} represents the 3D CA basis function and N is the filter size in the respective directions. c_{klm} denotes the transform coefficients whose values are obtained from the inverse transform of (1):

$$c_{klm} = \sum_{k=0}^{N_k-1} \sum_{k=0}^{N_l-1} \sum_{m=0}^{N_k-1} I(i,j,t) B_{ijtklm} \tag{2}$$

in which the bases B_{ijtklm} denote the inverse of A_{ijtklm}. 3D CA transform has the separability property, it corresponds to 2D CA transform in spatial domain followed by a 1D CA transform along the temporal axis. The 3D CA spatio-temporal transform model is shown in Fig. 2.

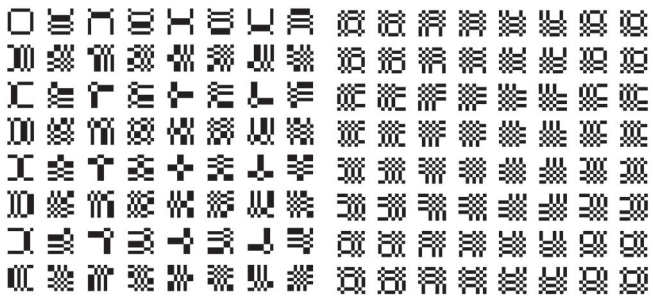


Fig. 3. Two different 2D dual-coefficient basis functions A''_{ijkl} using different CA gateways ($Rule = 112$ and 14): The white rectangular dots represent 1 while the black dots are 1. A''_{00kl} is the block at the extreme upper left corner, the top row represents $0 \leq j < 8, i = 0$.

The 3D CA basis function A_{ijtklm} can be generated in a variety of ways: Two-dimensional CA evolved over a specified time. In this approach, the basis function is given:

$$A_{ijtklm} = F(a_{ijt} a_{klm}) \tag{3}$$

where a denotes the states of the CA as they are evolved over time, F is a mapping function. Products of three orthogonal 1D CA transform filters:

$$A_{ijtklm} = A'_{ik} \times A'_{jl} \times A'_{tm} \tag{4}$$

where A'_{ik} denotes the 1D CA transform basis function and has been presented in our previous works [29], [30]. Products of orthogonal 2D (e.g., 2D, as used for images) and 1D CA transform filters. In this case

$$A_{ijtklm} = A''_{ijkl} \times A'_{tm} \tag{5}$$

where A' is 1D basis function while A'' denotes 2D bases and can be generated by a series of CA gateway keys ($e.g.$, Wolfram $Rule$: 14, 142; cell number: 8; initial configuration: 11010100, 0101110; boundary configuration: cyclic; basis function type: 2) and graphical display of these two rules are shown in Fig. 3.

CA machine is named multiprocessor machine because of the inherent property of the array of cellular automata. Each automaton is usually a simple processor capable of simple computational tasks. In a normal architecture, each of these processing nodes can interchange data only with its immediate neighbors and all processing nodes carry out the same computational operation. Although the operations available at each node are quite simple, the aggregated effect of many such nodes can exhibit complex behavior and can rapidly model quite complex dynamic systems. Computational speed is assured by the inherent parallelism of CA computing, in particular, CA can be considered as models of highly parallel and distributed computations in multiprocessor and distributed systems.

B. Watermark Robustness Improved by PRII

To improve the robustness of the watermark, in our work the watermark needs to be pre-processed by the computational integral imaging technique before embedding. This technique is divided into optical pickup part and computational integral imaging reconstruction (CIIR) part. In the pickup part, a set of elemental images is generated by the computer mapping algorithm instead of using a lenslet array. Fig. 4(a) displays the optical pickup device used in this experiment, and Fig. 4(b) describes the geometric principle of the 3D watermark pickup process used in the computational integral imaging system. Rays originating from the 3D watermark pass through each of lenslets and are captured as an EIA. The input watermark $O(y_0, z_0)$ is reproduced by the kth elemental lens and recorded at $y_{(c,k)}$ in the 2D image sensor. The pixel value $y_{(c,k)}$ of the kth elemental image of the object point $O(y_0, z_0)$ can be expressed as follows:

$$y_{(c,k)} = \frac{f_0 f_c (k\phi - y_0)}{f_d (z_0 - f_0)} - \frac{f_c k\phi}{f_d} \tag{6}$$

where f_0 denotes the focal length of each elemental lens, f_d represents the distance between the focal plane of the lenslet array and the camera lens, f_c is the focal length of the camera lens, and ϕ is the size of each elemental lens.

However, CIIR yields low-resolution images because there are intensity irregularities with a grid structure on the reconstructed image plane, which is a result of insufficient overlapping of the elemental images. To improve the limited resolution of CIIR, a super-resolution reconstruction technique is to be applied for integral imaging. The iterative back-projection super-resolution reconstruction technique as a potential and useful digital super-resolution technique was used in 3D image reconstruction [34], [38]. In this technique, a super-resolution image can be produced with an iterative of error estimation and an inference of a low-resolution image. Consider the imaging process of the observed image sequence g_k of a well-known method by $Irani$ and $Peleg$ [35]. The imaging model is modelled by:

$$g_k = \sigma(H_k * (T_k(f)) + \eta_k \tag{7}$$

where g_k corresponds to the observed image is the simulated sub-images. f is the reconstruct high resolution image. T_k denotes the 2D geometric transformation from f to g_k. H_k is a blurring operator, determined by the point spread function of the sensor. When lacking knowledge of the sensor properties it is assumed to be a Gaussian. η_k is the system noise, and "$*$" denotes the convolution operator.

Using Ji and $Fermuller$'s notation ([40]), the image formation process of (7) in the pixel domain can be modeled as

$$g_k = \sigma(H_k * F(T_k(\cdot))) + \eta_k \tag{8}$$

where F and g represent continuous signal and discrete signal, respectively. If we don't consider noise, the high resolution image f and low resolution image g_k can be expressed as

$$f = \sigma(F),\ g_k = [\sigma(H_k * F(T_k(\cdot))] \downarrow s \tag{9}$$

The relation between the high resolution image f and low resolution image g_k can be derived with the help of a difference ϵ_k ($\epsilon_k = T_k(x) - x$) [40]. Importing the difference ϵ_k into (9),

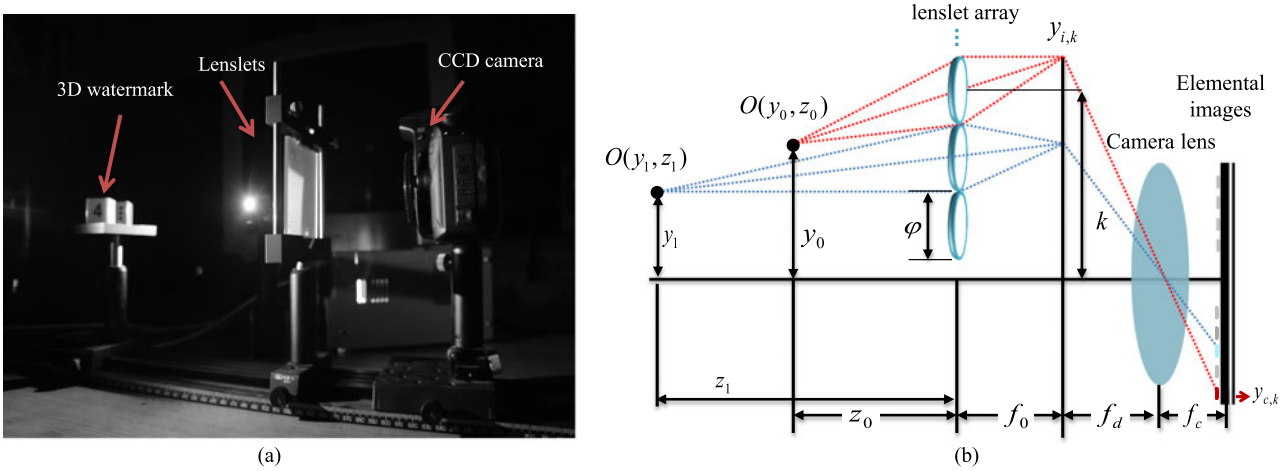


Fig. 4. (a) Optical pickup process, (b) geometric principle of the pickup process.

the low resolution image g_k also can be written as

$$\begin{aligned} g_k &= [\sigma(H_k * F(x+\epsilon_k)] \downarrow s \\ &= [\sigma(H_k * F(x) + \sigma(H_k * F(\epsilon_k))] \downarrow s \\ &= [\sigma(F(x) * H_k) + \sigma(F(\epsilon_k) * H_k)] \downarrow s \\ &= [f * H_k] \downarrow s + \epsilon_k [f * H_k'] \downarrow s \end{aligned} \tag{10}$$

Starting a initial guess $f^{(0)}$ for the high resolution image. The imaging process is simulated to obtain a set of low resolution observed images $g_k^{(n)}$ corresponding to the observed image $g_{(k)}$. The error function $e^{(n)}$ between the the observed and the simulated low resolution image is used to improve the high resolution image f, which is defined as

$$e^{(n)} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \|g_k - g_k^{(n)}\|_2^2} \tag{11}$$

where $g_k^{(n)}$ is the nth observed sub-image. If the calculated e^n is less than the given threshold value η and the iteration is terminated. Otherwise, the inferred super-resolution image is updated by

$$f^n = f^{n-1} + \frac{1}{K}\sum_{k=1}^{K}(T_k^{-1}[g_k - g_k^{(n-1)}]) \uparrow s] * p) \tag{12}$$

where T_k^{-1} is the inverse of T_k, p is the deblurring kernel, and $\uparrow s$ denotes up-sampling.

Iterative back-projection algorithm have been proven to be effective for super resolution reconstruction. However, in the iteration process shown in Fig. 5, the $(n-1)$th iteration inferred image exists ringing artifacts due to uniform amplification of the high-frequency components existing in the simulated sub-image and observed sub-image. Also, the deblurring process makes it very sensitive to the noise. Various regularization techniques have been presented to improve the robustness. However, these techniques either are very sensitive to the assumed noise model (Tikhonov regularization) or are computationally expensive [39]. In this work, we deal with these problems include

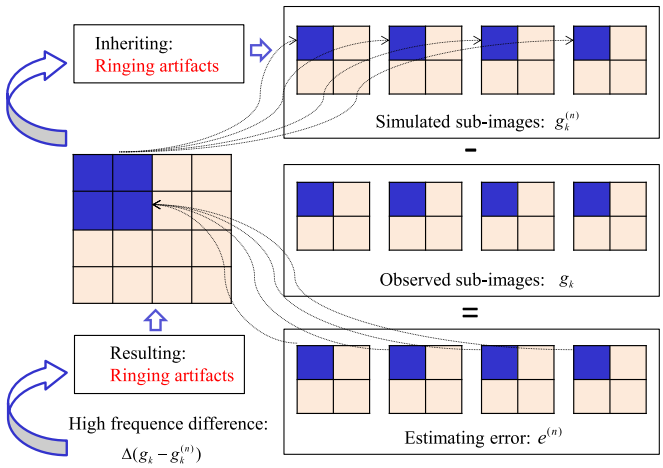


Fig. 5. A pictorial example of iterative back-projection method.

two folds. We first extract the high frequency difference to reduce the ringing artifacts. Then, we utilize the wavelet filter based perfect reconstruction filter bank to improve the robustness of the iterative algorithm. At first, as shown in Fig. 5, the conventional iterative back-projection will directly result in ringing artifacts existing in the nth iteration process. The ringing artifacts generally results from the high frequency difference between the simulated image model and the actual image data. To reduce the ringing artifacts of the inferred super-resolution image, we need to consider the high-frequency components in the update error process. In our experiment, the Laplace operator is employed to extract the high-frequency components of the simulated and observed sub-images. If Δ represents the Laplace operator, $\Delta(g_k - g_k^{(n)})$ proximately regards as high frequency difference components between the simulated and observed sub-images, thereby the $\Delta(g_k - g_k^{(n)})$ component should be employed into the simulated error that effectively removes ringing artifacts problems. The updated error can be revised as

$$e^{(n)} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \|(g_k - g_k^{(n)}) - \Delta(g_k - g_k^{(n)})\|_2^2} \tag{13}$$

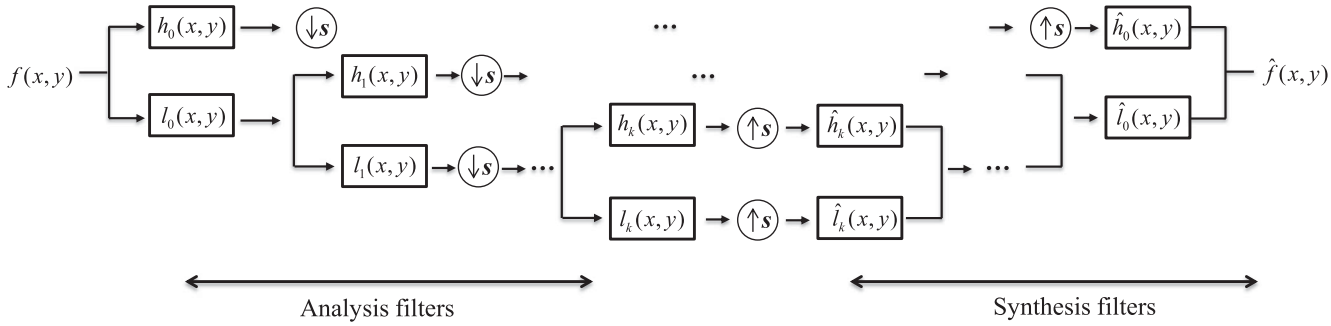


Fig. 6. Multiple-channel discrete wavelet transform on signal x.

Next, to further improve performance of iterative equation of (11), we import the wavelet perfect filter bank in it. The decomposition of signal into half band is essentially a wavelet transform on discrete sequences, since it splits the original space in two, and splits one of the resulting half space in two. This is shown in Fig. 6 using wavelet filters. The signal $f(x, y)$ through the wavelet filter bank, as shown in Fig. 6 can be expressed as

$$\hat{f} = \hat{h}_0 * [(h_0 * f) \downarrow s] \uparrow s + \hat{h}_1 * [(h_1 * f) \downarrow s] \uparrow s + \cdots + \hat{h}_k * [(h_k * f) \downarrow s] + \hat{\ell}_k * [(\ell_k * f) \downarrow s] \uparrow s \tag{14}$$

where $\{\hat{h}_k, \hat{\ell}_k\}$ and $\{h_k, \ell_k\}$ denote the high pass filter and low pass filter on the synthesis and analysis sides, respectively.

As shown in Fig. 6, for each low resolution signal g_k, using the format of (10), can be expressed in the form:

$$g_k = [\ell_k * f] \downarrow s + \epsilon_k[h_0 * f + h_1 * f + h_2 * f + \cdots + h_k * f] \downarrow s \tag{15}$$

From (15), we can export $[\ell_k * f] \downarrow s$ as follow:

$$[\ell_k * f] \downarrow s = g_k - \epsilon_k[h_0 * f + h_1 * f + h_2 * f + \cdots + h_k * f] \downarrow s \tag{16}$$

Combining use of (14) and (16), we can reconstruct the perfect signal $\hat{f}(x, y)$ from the low resolution signal g_k after nth iterative time by

$$\hat{f}^{(n+1)} = \hat{h}_0 * [(h_0 * \hat{f}^{(n)}) \downarrow s] \uparrow s + \cdots + \hat{h}_k * [(h_k * \hat{f}^{(n)}) \downarrow s] \uparrow s + \hat{\ell}_k * \frac{1}{N} \sum_{k=1}^{k=K} [g_k - \epsilon_k(h_0 * \hat{f}^{(n)} + \cdots + h_k * \hat{f}^{(n)}) \downarrow s] \uparrow s \tag{17}$$

Apply (14), we can simply rewrite (17) as follow:

$$\hat{f}^{(n+1)} = \hat{f}^{(n)} + \hat{\ell}_k * \frac{1}{K} \sum_{k=1}^{K} [g_k - (\ell_k * \hat{f}^{(n)} + \epsilon_k(h_0 * \hat{f}^{(n)} + h_1 * \hat{f}^{(n)} + \cdots + h_k * \hat{f}^{(n)})) \downarrow s] \uparrow s \tag{18}$$

Such a filter bank is named perfect filter bank if $f = \hat{f}$ for all input signal f. For two-channel filter bank, in that case, (18) is rewritten as

$$\hat{f}^{(n+1)} = \hat{f}^{(n)} + \hat{\ell} * \frac{1}{K} \sum_{k=1}^{K} [g_k - (\ell * \hat{f}^{(n)} + \epsilon_k(h * \hat{f}^{(n)})) \downarrow s] \uparrow s \tag{19}$$

Comparing with the previous presented iterative equation of (12), our method does not need to consider the complex geometric transform process between g_k and $\hat{f}$. The deblurring kernel p is also replaced by $\hat{\ell}$. For two-channel filter bank, the analysis filters $\{h, \ell\}$ and the synthesis filters $\{\hat{h}, \hat{\ell}\}$ of a perfect reconstruction have to satisfy the following condition (see more details in [40]–[42]):

$$\hat{h}(z) = \ell(-z), \hat{\ell}(z) = -h(-z), \tag{20}$$

$$\|\hat{h} * h\| = \|\hat{\ell} * \ell\| = \frac{1}{2} \tag{21}$$

The deblurring kernel p in (12) should be a good approximation for the inverse of ℓ. However, the deblurring kernel $\hat{\ell}$ in (19) only needs to satisfy the condition $\|\hat{\ell} * \ell\| = \frac{1}{2}$, that means that it possible to design $\hat{\ell}$ that has much smaller norm than p. This indicates that this iterative algorithm is more robust than conventional methods and much better resistance to noise propagation.

Apply the update error of (13), we can rewrite the updated iterative equation of (19) in the form

$$\hat{f}^{(n+1)} = \hat{f}^{(n)} + \hat{\ell} * \frac{1}{K} \sum_{k=1}^{K} [(g_k - (\ell * \hat{f}^{(n)} + \epsilon_k(h * \hat{f}^{(n)})) - \Delta(g_k - (\ell * \hat{f}^{(n)} + \epsilon_k(h * \hat{f}^{(n)})))) \downarrow s] \uparrow s \tag{22}$$

At last, about the low resolution images used in the reconstruction process, unlike the previous CIIR-based iterative algorithm in [34], [38], the elemental images were directly used as the low resolution images for iteration. In our work, to increase the quality of elemental images, a transformation algorithm between the EIA and the sub-image array is utilized and the transformed results will be found in Fig. 12(b) and (e). The sub-image array can be generated from the elemental images by the following transformational process: an EIA is composed of x and y elemental images, the EIA with the size and the x-line and y-column elemental image (EI) can be

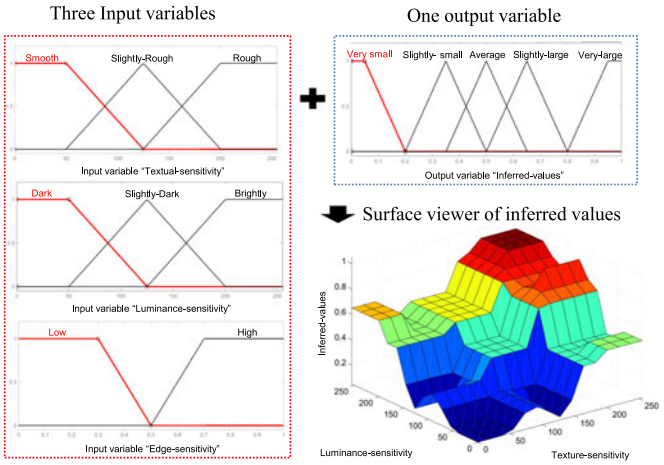


Fig. 7. Input/output membership functions: left denotes the three input variances Luminance-sensitivity, Texture-sensitivity, Edge-sensitivity, top-right denotes the output variance Inferred-values, bottom-right denotes the surface viewer of the inferred values.

calculated by

$$EI_{x,y}(m,n) = EIA(xM+m, yN+n) \tag{23}$$

here, $M \times N$ represents the size of each of elemental images, $m = 1, 2, \ldots, M.n = 1, 2, \ldots, N$. Hereby, the m-line n-column sub-image (SI) can be obtained from

$$SI_{m,n}(x,y) = EIA(xM+m, yN+n) \tag{24}$$

From (24), we can see that the new generated sub-image array contains sub-images and each image with the size of $x \times y$.

C. Embedding Parameters Inferred by Fuzzy Inference System

We first infer the embedding parameters α, to enhance data-hiding capacity and the quality of the watermarked video sequence by using fuzzy inference system (FIS). Generally, uniform regions of image are sensitive to the embedded watermark so, only a small amount of data can be added in these areas. The human visual system (HVS) has been considered with several phenomena that permit to adjust the pixel values to elude perception [43], [44]. The fuzzy interference system is used here to adapt the HVS different properties. In this work, we are considering texture, luminance and edge sensitivity of video frames, so that embedding the watermark information in these features makes the watermark imperceptible. Fig. 7 shows the input/output membership functions and the surface viewer of the inferred values by Matlab graphic user interface editor.

As for gray-scale images, the mean values of the blocks can be used as measuring the luminance values of each block. Brighter areas are less sensitive to the visibility than to darker region. The luminance values of the k-block can be determined as follows:

$$L_s^k(x,y) = \frac{1}{M_k \times Nk} \sum_{x=0}^{M_k-1} \sum_{y=0}^{N_k-1} c_k(x,y) \tag{25}$$

where $c_k(x,y)$ represents the pixel values of the kth block of the CA transform coefficient, and $M_k \times N_k$ denotes the size of each block.

Besides, the human eyes are less sensitive to highly textured regions; therefore the bigger embedding parameters appear in the highest texture areas. The textual sensitivity of images can be calculated by

$$T_s^k(x,y) = \sum_{x=0}^{M_k-1} \sum_{y=0}^{N_k-1} |c_k(x,y) - L_s^k(x,y)| \tag{26}$$

To make lower visibility, a watermark is added in higher edge areas [43]. In this work, the Canny algorithm is selected to calculate the edge sensitivity. The centroid functions in [45] are introduced to defuzzify the inference results.

D. Watermarking Embedding

1) The watermark is preprocessed by the PRII before embedding and the sub-image array is obtained.
2) To further improve the security of watermark, the sub-image array is scrambled by 1D CA sequence, and the scrambled data can be generated according to:

$$c_k = \sum_{i=0}^{N-1} f_i B_{ik}, \tag{27}$$

$$B_{ik} = \frac{A_{ik}}{\sum_{j=0}^{N-1} A_{jk}^2} \tag{28}$$

where f_i represents the 1D watermark sequence and A_{ik} is the 1D CA basis function.
3) The uncompressed raw video sequence is separated into a group of frames.
4) Make pseudo-3D CA transform for each frame and further obtain the corresponding 3D CA transform coefficients. The 3D CA transform coefficients can be generated according to the following equation:

$$c(k,l,m) = \sum_{i=0}^{N_k-1} \sum_{i=0}^{N_l-1} \sum_{i=0}^{N_m-1} f_{ijt} B_{ijtklm} \tag{29}$$

where B_{ijtklm} is the inverse of the CA basis function A_{ijtklm}.
5) The CA transform coefficients can be further transformed and the goal is to find optimized coefficients for watermark embedding.
6) The optimized CA transform coefficients are divided into v blocks, each block is evaluated by the designed FIS to infer the embedding weight function $\alpha^{(v)}$ for all the pixels in the block, and then a criteria is applied to infer the best weight factors to be embedded.
7) The scrambled data is embedded into the candidate CA transform coefficients by using:

$$\hat{c}(k,l,m) = c^v(k,l,m) + W^{(v)}(k,l,m) \times \alpha^v(k,l,m) \tag{30}$$

where $c^v(k,l,m)$ represents the vth block of the 3D CA transform coefficients from the frame $I(k,l)$ at temporal

axis $t = m$, $W^{(v)}(k, l, m)$ denotes the vth block of scrambling watermark data at the time $t = m$, $\alpha^v(k, l, m)$ denotes the embedding parameter of the vth block referred by the designed FIS and the calculation process can be found in Section II-C.

E. Watermarking Extraction

1) The watermarked video sequence is first separated into a group of frames.
2) Make pseudo 3D CA transform for each frame and obtain the corresponding 3D CA coefficients.
3) In this work, the watermark can be detected using two ways: blind and non-blind. Non-blind watermark detection is that the receiver needs the original image information to extract watermark data. Due to the information of gray watermark is more huge than the binary watermark, non-blind extraction method is best for gray watermark:

$$\hat{W}(k, l, m) = \frac{\hat{c}^v(k, l, m) - c^v(k, l, m)}{\alpha^v(k, l, m)} \tag{31}$$

where $\hat{W}(k, l, m)$ represents the extracted watermark with pixel values from 0 to 255.
4) The extracted watermark data is inversely scrambled by 1D CA and the decrypted sub-image array is obtained.
5) The robust watermark is reconstructed from the recovered sub-image array using the PRII reconstruction technique (see the reconstruction algorithm in (22)).
6) Blind watermark detection is accomplished without referring the original image. However, it is more suits to the binary watermark. For selected candidate watermarked CA transform coefficients $\hat{C}$ and the scrambled watermark data W, we calculate the correlation ρ between them, that is

$$\rho = \frac{\sum_{i=0}^{N-1}(\hat{C}_i - \bar{\hat{C}})(W_i - \overline{W})}{\sqrt{\sum_{i=0}^{n-1}(\hat{C}_i - \bar{\hat{C}})^2}\sqrt{\sum_{i=0}^{n-1}(W_i - \overline{W})^2}} \tag{32}$$

where $\overline{x}$ denotes the mean values of x. The watermark detection is performed by comparing the correlation to an adaptive threshold value κ.

$$W = \begin{cases} 1, & \text{if } \rho > \kappa \\ 0, & \text{if } \rho \leq \kappa \end{cases}$$

where the adaptive threshold κ depends on the mean values of ρ. $\kappa = \beta \times \overline{\rho}, \beta = (2, 2.5)$, therefore the embedded data can be retrieved lossless.
7) The extracted watermark data W needs to be further inversely transformed by (24), and the binary-scaled sub-image array is obtained.
8) The final watermark can be reconstructed from the recovered sub-images by the PRII reconstruction algorithm (see (22)).

III. Experimental Results

To show the effectiveness of this work, numerical simulations have been performed under the environment of software Matlab 8.1 on a private computer (Intel Core I7-3770 3.9 GHz CPU with 8 GB RAM). The PRII technique is implemented by use of a virtual computer synthetic pinhole array that is composed of 30×30 pinholes. The diameter of each pinhole is set to 1.08 mm and the gap d between the pickup device and the virtual pinhole array is 3 mm. The pickup distance is set to $l = 36$ mm. The CA gateway keys for transform using 8-bit CA, initial configuration 01001101, Wolfram $Rule = 14$. The performance evaluated on three standard video sequences: '$Akiyo$', '$Foreman$' and '$Suzie$' with uncompressed AVI format, of size of 480×360 and frame rate of 30 fps. Each video consists of 360 frames. Some frames of these three video sequences are selected and presented in Fig. 8. Watermark '$SCU - IID$' and '$dice$' with the size of 120×90 are used as the watermarks shown in Fig. 8 (d) and (e), respectively. The iteration of the PRII system is set to $n = 50$ and the given threshold value of error η is set to 0.0012. In our work, the extracted watermark data from the watermarked video sequence belongs to sub-images, which is not the final watermark. The final watermark can be reconstructed from these extracted sub-images by utilizing the PRII reconstruction technique.



Fig. 8. Original frames from video sequence (uncompressed AVI format, 480×360 pixels, 30 fps): (a) '$Akiyo$', (b) '$Foreman$', (c) '$Suzie$', (d) watermark logo '$SCU - IID$', (e) '$dice$'.

The imperceptibility and robustness are used to measure the watermarking method performance. For imperceptibility, the peak signal-to-noise ratio $(PSNR)$ and structural similarity index $(SSIM)$ are used for measuring watermarked video quality. About the description of $PSNR$ can be found in [30]. $SSIM$ is considered to be correlated with the quality perception of HVS and designed by modeling any image distortion as a combination of three factors that are distortion of structure $s(x, y)$, luminance distortion $l(x, y)$ and contrast distortion $c(x, y)$. The $SSIM$ is defined as:

$$SSIM(x, y) = s(x, y)l(x, y)c(x, y) \tag{33}$$

where the definition of these three factors $s(x, y)$, $l(x, y)$, and $c(x, y)$ can be found in [46]. The overall image quality can be evaluated by the mean of $SSIM$, and the positive value of $SSIM$ is in the range between 0 to 1, where 1 indicates no image-quality distortion. For robustness, $SSIM$ and bit correct ratio (BCR) measures the similarity between the original and extracted watermarks.

Fig. 9 shows the proposed watermarked video sequences with the help of fuzzy inferred alpha, and the corresponding $SSIM$ index maps are shown in Fig. 10. The average $SSIM$ and $PSNR$ values of these three watermarked video sequences are recorded in Table I. Results show that no matter via human

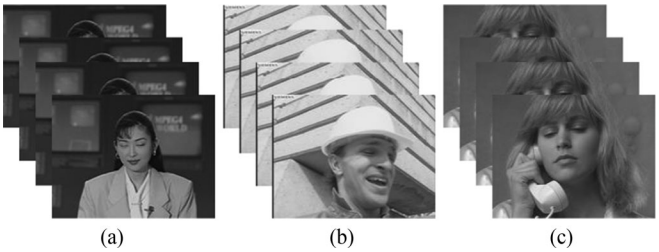


Fig. 9. (a) Watermarked video sequences: (a) $Akiyo$, (b) $Foreman$, (c) $Suzie$.

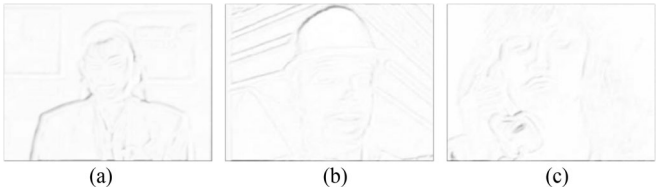


Fig. 10. Mean $SSIM$ value map of the one selected frame of three video sequences: (a) $Akiyo$, (b) $Foreman$, (c) $Suzie$.

TABLE I
AVERAGE $SSIM$ AND $PSNR$ VALUES OF THREE SELECTED VIDEO FRAMES OF FIG. 9

$SSIM$			
Watermarks	$Akiyo$	$Foreman$	$Suzie$
$SCU-IID$	0.9863	0.9815	0.9837
dice	0.9762	0.9687	0.9759
$PSNR(dB)$			
Watermarks	$Akiyo$	$Foreman$	$Suzie$
$SCU-IID$	56.13	54.37	55.96
dice	54.33	53.29	53.97

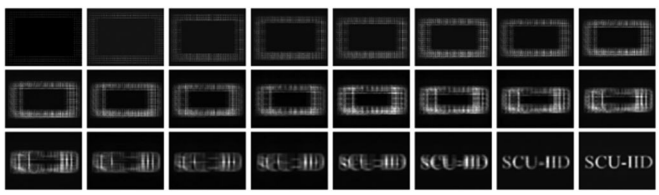


Fig. 11. Reconstruction process of the elemental images against occlusion attack.

observation or quantitative analysis, the transparency of our proposed fuzzy-based watermarking method is satisfied.

In this work, the reason of utilizing the PRII to preprocess watermark it is that the memory distribution properties of the elemental images which greatly increase robustness. Fig. 11 shows an example of the reconstruction process of an occlusion-attacked EIA. From these results presented in Fig. 11, we can see that even though most of the elemental images are occluded, the watermark is able to be reconstructed successfully. In this work, to further improve the image-quality of the CIIR technique, the improved iterative PRII technique is employed to reconstruct the

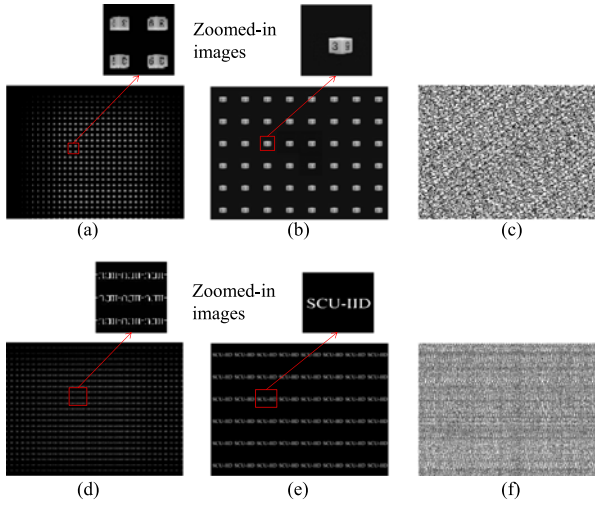


Fig. 12. (a) EIA of '$dice$', (b) transformed sub-image array, (c) scrambled noise-style image, (d) EIA of '$SCU-IID$', (e) transformed sub-image array, (f) scrambled noise-style watermark.

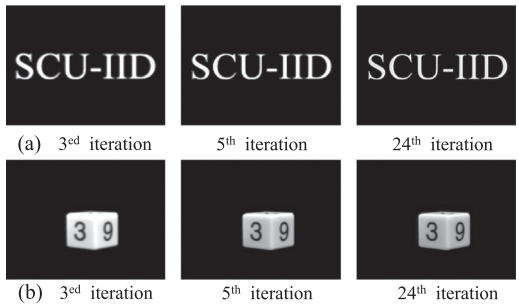


Fig. 13. (a) Extracted watermark logo '$SCU-IID$' and (b) '$dice$' after different iterative times.

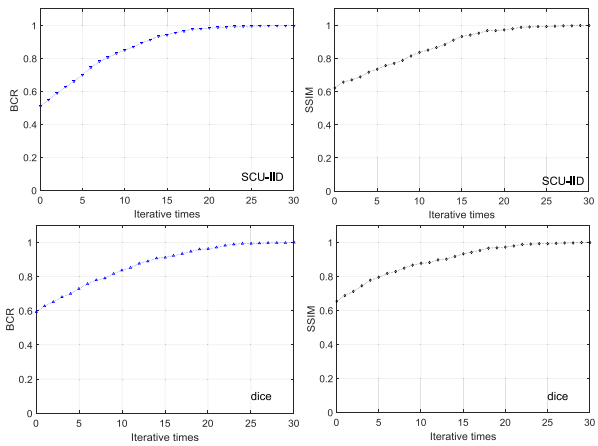


Fig. 14. (a) BCR and $SSIM$ values of the extracted watermark '$SCU-IID$' and '$dice$' after different iterative times.

elemental images, and the reconstruction results of our method after different iterative times will be shown below.

Fig. 12 shows the results of the watermark with the pre-processing algorithm. Fig. 12(a) shows the 30×30 elemental images of dice. Fig. 12(b) shows the transformed 8×6 sub-images by (20). Fig. 12(c) shows the scrambled watermark data, which as the final watermark data is embedded

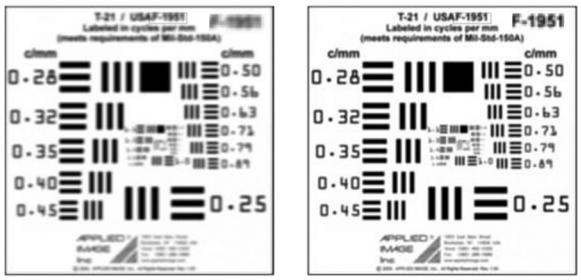


Fig. 15. Reconstructed resolution target 1951USAF resolution chart after 24 iterations: (a) IBP-based CII method in [34], (b) proposed method.



Fig. 16. (a) Reconstructed watermark plane image '$SCU-IID$' with the proposed method, (b) reconstructed watermark plane image '$SCU-IID$' with the IBP method in [34].

into CA transform coefficients. Fig. 12(d)–(f) shows another pre-processing results for '$SCU-IID$'. From the scrambled results in Fig. 12(c) and (f), we can infer that the watermark data is well encrypted.

Fig. 13(a) and (b) shows the reconstructed '$SCU-IID$' and '$dice$' using the PRII technique after 3ed, 5th, and 24th iterative times, respectively. The results are presented in Fig. 13 show that the watermark can be clearly reconstructed after the iterative time of 24. Meanwhile, the corresponding BCR and $SSIM$ curve-graphs of '$SCU-IID$' and '$dice$' with different iterative times are shown in Fig. 13. The results in Fig. 14 show that the curves of BCR and $PSNR$ become horizontal gradually when the iterative time reaches to 24 times.

To validate the proposed algorithm, we perform a simulation experiment to compare the performance of the proposed PRII-based method with the IBP-based CII method [34]. In this simulation, a simulated resolution target 1951USAF resolution chart with the size of 256×256 is used. Fig. 15(a) shows the reconstructed resolution target 1951USAF resolution chart after 24 iterations with the method of [34], Fig. 15(b) the proposed reconstructed resolution target 1951USAF resolution chart after 24 iterations. From the content located in the red box, we can know that our method provides better resolution than the IBP based method. Fig. 16 shows another example with the watermark plane image '$SCU-IID$', Fig. 16(a) and (b) shows the reconstructed plane images with our method and the method in [34]. Fig. 17(a) shows the proposed reconstructed watermarks '$dice$' after the 3rd, 5th and 24th iteration, respectively. Fig. 17(b) shows the reconstructed images '$dice$' with the $A.Stern$ and $B.Javidi$ method in [34] after the 3rd, 5th and 24th iteration, respectively. Compared with the results showed

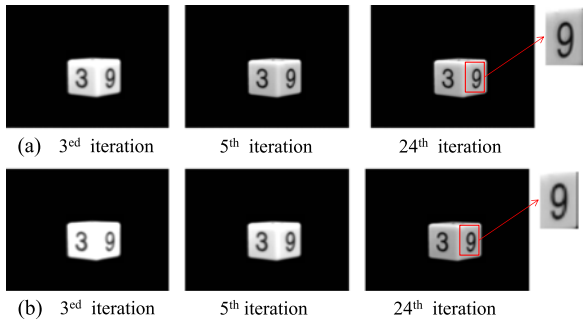


Fig. 17. (a) Reconstructed watermark plane images with the proposed method, (b) reconstructed watermarks with the IBP method in [34].

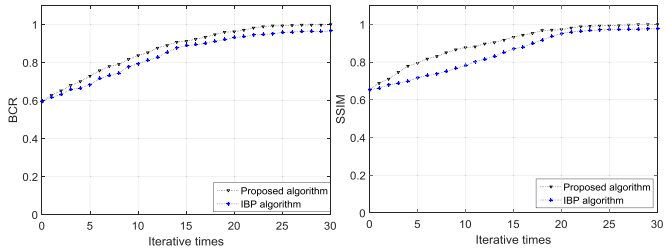


Fig. 18. BCR and $SSIM$ curves of increasing iterative times with the proposed scheme and A. Stern and B. Javidi algorithm in [34].

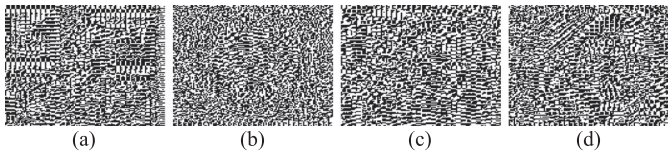


Fig. 19. Extracted watermarks with different CA gateway keys: (a) and (b) extracted SCU-IID and dice with Wolfram $Rule = 15$; (c) and (d) extracted SCU-IID and dice with initial configuration: 11001101. Correct CA gateway keys: Wolfram $Rule = 14$; initial configuration 01001101 used in this experiment.

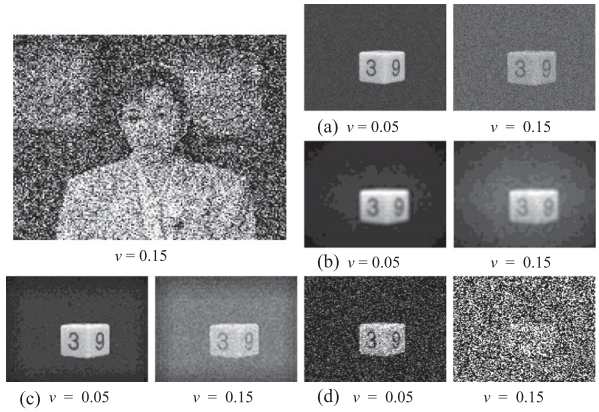


Fig. 20. Reconstructed watermarks from Gaussian noise attack with three methods: (a) proposed method, (b) method in [30], (c)method in [29], (d) gray watermark based method in [22].

in Fig. 17, we can infer that the PRII technique has effectively removed the ringing artifacts problem in the IBP super-resolution technique.

The convergence BCR and $SSIM$ curves of two super-resolution image reconstruction algorithms are shown in Fig. 18. The convergence speed of the optimized method is obviously to

TABLE II
AVERAGE $SSIM$ AND BCR VALUES AGAINST GAUSSIAN NOISE ATTACK

	Gaussian noise	$SSIM$			BCR		
Watermarking methods	"V"	$Akiyo$	$Foreman$	$Suzie$	$Akiyo$	$Foreman$	$Suzie$
Proposed	0.05	0.9362	0.9287	0.9359	0.9263	0.9261	0.9113
	0.15	0.8762	0.8395	0.8627	0.8346	0.8493	0.8458
Method in [22]	0.05	0.8651	0.8532	0.8637	0.8413	0.8365	0.8397
	0.15	0.5953	0.5152	0.5012	0.5844	0.5762	0.5632
Method in [29]	0.05	0.9136	0.9027	0.9133	0.9016	0.8993	0.8910
	0.15	0.8601	0.8190	0.8419	0.8017	0.8143	0.8226
Method in [30]	0.05	0.8632	0.8517	0.8553	0.8617	0.8567	0.8602
	0.15	0.7513	0.7462	0.7512	0.7366	0.7213	0.7369

TABLE III
AVERAGE $SSIM$ AND BCR VALUES AGAINST SALT & PEPPER NOISE ATTACK

	Salt & pepper	$SSIM$			BCR		
Watermarking method	"D"	$Akiyo$	$Foreman$	$Suzie$	$Akiyo$	$Foreman$	$Suzie$
Proposed	0.05	0.9217	0.9305	0.9214	0.9003	0.9327	0.9258
	0.10	0.8543	0.8216	0.8327	0.8016	0.8359	0.8236
Method in [22]	0.05	0.8319	0.8137	0.8675	0.7965	0.8163	0.7966
	0.10	0.5537	0.5854	0.5917	0.5532	0.5855	0.5437
Method in [29]	0.05	0.9079	0.9123	0.9028	0.8795	0.9137	0.8959
	0.10	0.8367	0.8013	0.8102	0.7901	0.8009	0.7937
Method in [30]	0.05	0.8517	0.8436	0.8331	0.8426	0.8502	0.8400
	0.10	0.7762	0.7651	0.7315	0.7517	0.7412	0.7533

TABLE IV
AVERAGE $SSIM$ AND BCR VALUES AGAINST FRAME CLIPPING ATTACK

	Frame clipping	$SSIM$			BCR		
Watermarking method	"S"	$Akiyo$	$Foreman$	$Suzie$	$Akiyo$	$Foreman$	$Suzie$
Proposed	40%	0.9762	0.9687	0.9759	0.9637	0.9739	0.9630
	60%	0.8878	0.8639	0.8753	0.8631	0.8657	0.8738
Method in [22]	40%	0.7315	0.7261	0.7749	0.7187	0.7.36	0.7237
	60%	0.5869	0.5766	0.5369	0.5322	0.5137	0.5348
Method in [29]	40%	0.9436	0.9325	0.9431	0.9368	0.9401	0.9228
	60%	0.8126	0.8327	0.8007	0.7958	0.8138	0.8265
Method in [30]	40%	0.7862	0.7637	0.7542	0.7622	0.7713	0.7753
	60%	0.6512	0.6631	0.6712	0.6632	0.6713	0.6516

TABLE V
AVERAGE $SSIM$ AND BCR VALUES UNDER JPEG COMPRESSION ATTACK

	JPEG compression	$SSIM$			BCR		
Watermarking method	"Q"	$Akiyo$	$Foreman$	$Suzie$	$Akiyo$	$Foreman$	$Suzie$
Proposed	50%	0.9482	0.9327	0.9436	0.9428	0.9355	0.9367
	10%	0.8862	0.8713	0.8901	0.9017	0.8934	0.8806
Method in [22]	50%	0.8762	0.8963	0.8755	0.8563	0.8322	0.8156
	10%	0.7563	0.7462	0.7813	0.7468	0.7352	0.7219
Method in [29]	50%	0.9136	0.9021	0.9106	0.9163	0.9038	0.9057
	10%	0.8433	0.8369	0.8437	0.8530	0.8501	0.8491
Method in [30]	50%	0.8613	0.8852	0.8743	0.8632	0.8612	0.8333
	10%	0.7927	0.8012	0.7855	0.7917	0.8013	0.7813

TABLE VI
AVERAGE $SSIM$ AND BCR VALUES UNDER MEDIAN FILTER ATTACK

Watermarking method	Median filter "W"	$SSIM$			BCR		
		$Akiyo$	$Foreman$	$Suzie$	$Akiyo$	$Foreman$	$Suzie$
Proposed	3×3	0.9537	0.9459	0.9436	0.9651	0.9538	0.9427
	5×5	0.8536	0.8437	0.8566	0.8364	0.8351	0.8260
Method in [22]	3×3	0.8217	0.8143	0.8328	0.8137	0.8256	0.8013
	5×5	0.6713	0.6638	0.6219	0.6349	0.6658	0.6712
Method in [29]	3×3	0.9368	0.9220	0.9157	0.9327	0.9287	0.9159
	5×5	0.8146	0.8257	8.8336	0.8008	0.8029	0.7937
Method in [30]	3×3	0.8752	0.8622	0.8503	0.8453	0.8537	0.8612
	5×5	0.7315	0.7207	0.7462	0.7503	0.7433	0.7612

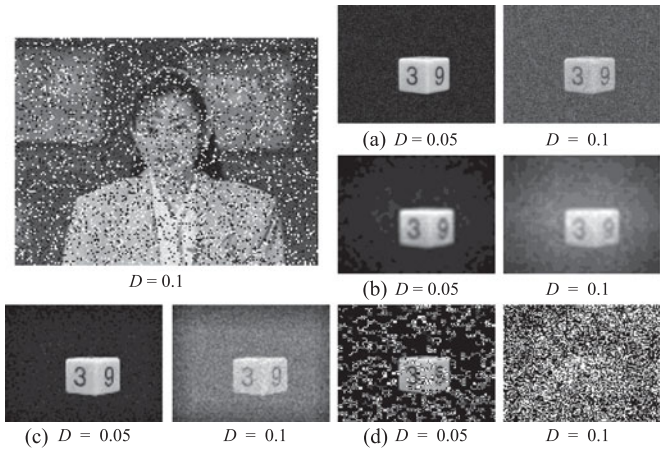


Fig. 21. Reconstructed watermarks from salt and pepper noise attack with three methods: (a) proposed method, (b) method in [30], (c) method in [29], (d) gray watermark based method in [22].

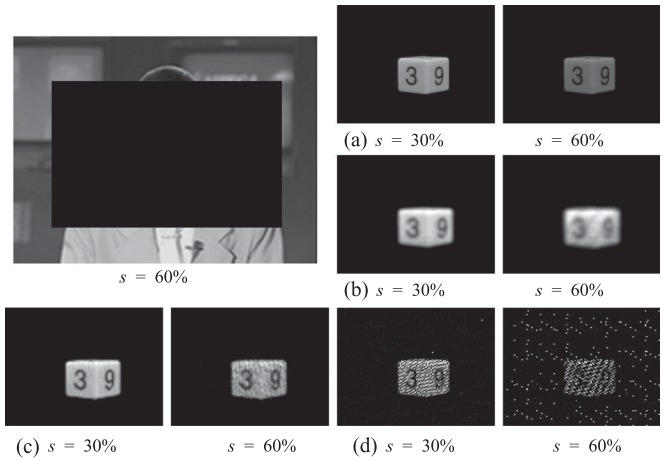


Fig. 22. Reconstructed watermarks from frame clipping attack with three methods: (a) proposed method, (b) method in [30], (c) method in [29], (d) gray watermark based method in [22].

be faster than the conventional IBP method. And it also means that the optimized method effectively avoid interference caused by the ringing artifacts problem in $A.Stern$ and $B.Javidi$ [34].

CA can provide multiple transform planes for watermark embedding. Using a series of CA gateway keys can generate many CA basis functions (see Fig. 3), which can provide multiple transform planes for watermark embedding. Fig. 19 shows the reconstructed watermarks using incorrect CA gateway keys. The simulation results show that CA gateway keys not only can provide multiple transform planes for watermark embedding, but also increase the security of watermarking system. Compared to the transform-based watermarking works, our method has been increased $(2^8)_{\text{rules}} \times (2^8)_{\text{configurations}}$ security keys at least.

In addition, an effective video watermarking system should need to tolerate a certain amount of attacks, such as Gaussian noise, salt & pepper, frame clipping attack, lossy compression attack and filter attack. In this experiment, the gray watermark '$dice$' is used to evaluate robustness. For comparative purposes, the following similar typical methods from the literature are implemented: The first method is the computational integral imaging based watermarking, in which a 3D EIA is embedded into the 2D CAT transform domain in [30], in their approach, the watermark is recovered by CIIR. In [29], the authors proposed a multiple watermark based watermarking method, in which the computer-generated integral imaging algorithm was used for recording and reconstructing the watermark. The third method is gray watermark based watermarking method [22], in which a gray watermark is embedded the DCT coefficients. To improve the robustness of their proposed approach, the gray watermark was first decomposed into a series of binary bitplanes. However, in our experiment, a gray watermark is first divided into a series of memory-distributed elemental images by PRII. We also analyze the performance of our proposed memory-distributed watermark with gray watermark based method in $Niu\ et\ al.$ method [22].

In our experiment, the video sequence is distorted by Gaussian noise, salt & pepper noise, Frame clipping, JPEG compression, and median filtering. After the attacks, each watermark is extracted and is compared with the original watermark in terms of $SSIM$ and BCR. Fig. 20 shows the results after those of Gaussian noise (mean $= 0$, variance $(V) = 0.05$ and $V = 0.15$) compared our proposed algorithm with the method in [22], the method in [29], and the method in [30].

The average values $SSIM$ and BCR of the watermarked $Akiyo$, $Foreman$ and $Suzie$ with the $V = 0.05$ and $V = 0.15$ are recorded in Table II. The results clearly demonstrate that the proposed method consistently outperforms the method in

TABLE VII
AVERAGE TIME OF SEQUENCE '$Akiyo$' TAKEN BY DIFFERENT METHODS

Time-consume	Proposed	[22]	[29]	[34]
Embedding time ($sec.$)	47.23	65.37	55.26	56.16
Extraction time ($sec.$)	31.63	50.61	41.13	43.16

TABLE VIII
AVERAGE TIME OF VIDEO '$Foreman$' TAKEN BY DIFFERENT METHODS

Time-consume	Proposed	[22]	[29]	[34]
Embedding time ($sec.$)	45.86	68.17	54.38	56.16
Extraction time ($sec.$)	33.61	52.36	40.03	44.16

TABLE IX
AVERAGE TIME OF VIDEO '$Suzie$' TAKEN BY DIFFERENT METHODS

Time-consume	Proposed	[22]	[29]	[34]
Embedding time ($sec.$)	48.27	70.03	58.52	61.37
Extraction time ($sec.$)	34.08	55.69	44.25	48.36

[22], the method in [29], and the method in [30]. Obviously, the higher the $SSIM$ and BCR value, the better the degree of robustness.

The reason of the robustness of our work is super to [29] and [30] is that the perfect-reconstruction integral imaging effect address the blur noise existing previous CIIR-based watermarking works. Comparing with [22], the memory-distributed property of elemental images greatly improve the robustness of watermark.

Except Gaussian noise attack, the $SSIM$ and BCR values against salt and pepper noise, frame clipping, JPEG compression and median filter attack are recorded in Tables III–VI. Figs. 21 and 22 show the results of salt pepper noise and frame clipping attacks, respectively. From these results, we can know that the $SSIM$ and BCR values evidently better than that for the methods of in [22], [29], and [30].

To estimate the time complexity of CA transform-based video watermarking, we evaluate time-consume for 450 frames of each of video sequence '$Akiyo$', '$Foreman$' and '$Suzie$' with four different watermarking methods. Here, in order to effectively compare the time-consume of our proposed PRII method with the IBP-based CII method in term of watermarking. We only adopt the IBP reconstruction algorithm from [34], the watermarking embedding and extraction algorithms used are the same with our method. The estimated time is recorded in Tables VII–IX. The video sequence can be carried out in parallel on the CA multi-processor machine, thus, the proposed watermarking method which is a potential candidate for implementing video watermarking on a real-time.

IV. CONCLUSION

In this paper, we presented a new invisible robust video watermarking system using the 3D CA transform. To the best of our knowledge, this is the first video watermarking architecture utilizes 3D CA transform. Unlike the previous DWT, DCT, and DFT based watermarking methods, the proposed method provides multiple transform planes for watermark embedding, thus increased watermark security dramatically. Furthermore, we employed the PRII technique to deal with the watermark. The gain in PRII could effectively employ distributed memory to combat data loss attack. Unlike most of the existing integral imaging techniques, the proposed PRII technique can effectively resolve the ringing artifacts problem. Also, in the course of our investigation, it was shown that the proposed system achieves higher image quality and data-hiding capacity with the benefit of the fuzzy system. Finally, the performance of the novel watermarking system is presented where the robustness against typical watermark attacks, such as Gaussian, salt pepper noise, frame clipping, JPEG compression, median filtering, are highlighted. Also, it has been demonstrated that improvement in embedding time-consume can be achieved using the CA multiprocessor machine.

REFERENCES

[1] A. Stern and B. Javidi, "Three-dimensional image sensing, visualization, and processing using integral imaging," *Proc. IEEE*, vol. 94, no. 3, pp. 591–607, Mar. 2006.

[2] J.-H. Park, K. Hong, and B. Lee, "Recent progress in three-dimensional information processing based on integral imaging," *Appl. Opt.*, vol. 48, no. 34, pp. H77–H94, 2009.

[3] X. Xiao, B. Javidi, M. Martinez-Corral, and A. Stern, "Advances in three-dimensional integral imaging: Sensing, display, and applications [Invited]," *Appl. Opt.*, vol. 52, no. 4, pp. 546–560, 2013.

[4] P. W. Wong and N. Memon, "Secret and public key image watermarking schemes for image authentication and ownership verification," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1593–1601, Oct. 2001.

[5] M. Asikuzzaman, M. J. Alam, A. J. Lambert, and M. R. Pickering, "Robust DT CWT-based DIBR 3D video watermarking using chrominance embedding," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1733–1748, Sep. 2016.

[6] M. Chen, Z. Chen, X. Zeng, and Z. Xiong, "Model order selection in reversible image watermarking," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 3, pp. 592–604, Jun. 2010.

[7] H. V. Zhao and K. J. R. Liu, "Impact of social network structure on multimedia fingerprinting misbehavior detection and identification," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 4, pp. 687–703, Aug. 2010.

[8] C. Li, "Cracking a hierarchical chaotic image encryption algorithm based on permutation," *Signal Process.*, vol. 118, pp. 203–210, 2016.

[9] S. Li, C. Li, K.-T. Lo, and G. Chen, "Cryptanalysis of an image scrambling scheme without bandwidth expansion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 3, pp. 338–349, Mar. 2008.

[10] C. Li, Y. Liu, T. Xie, and M. Z. Q. Chen, "Breaking a novel image encryption scheme based on improved hyperchaotic sequences," *Nonlinear Dyn.*, vol. 73, no. 3, pp. 2083–2089, 2013.

[11] S. Li, C. Li, K.-T. Lo, and G. Chen, "Cryptanalyzing an encryption scheme based on blind source separation," *IEEE Trans. Circuits Syst.–I*, vol. 55, no. 4, pp. 1055–1063, May 2008.

[12] L. Luo, Z. Chen, M. Chen, X. Zeng, and Z. Xiong, "Reversible image watermarking using interpolation technique," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 1, pp. 187–193, Mar. 2010.

[13] M. Fallahpour, S. Shirmohammadi, M. Semsarzadeh, and J. Zhao, "Tampering detection in compressed digital video using watermarking," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 5, pp. 1057–1072, May 2014.

[14] P. W. Chan, M. R. Lyu, and R. T. Chin, "A novel scheme for hybrid digital video watermarking: Approach, evaluation and experimentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 12, pp. 1638–1649, Dec. 2005.

[15] L. E. Coria, M. R. Pickering, P. Nasiopoulos, and R. K. Ward, "A video watermarking scheme based on the dual-tree complex wavelet transform," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 3, pp. 466–474, Sep. 2008.

[16] H.-Y. Huang, C.-H. Yang, and W.-H. Hsu, "A video watermarking technique based on pseudo 3-D DCT and quantization index modulation," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 4, pp. 625–637, Dec. 2010.
[17] S. D. Lin and C.-F. Chen, "A robust DCT-based watermarking for copyright protection," *IEEE Trans. Consum. Electron.*, vol. 46, no. 3, pp. 415–421, Aug. 2000.
[18] J. C. Patra, J. E. Phua, and C. Bornand, "A novel DCT domain CRT-based watermarking scheme for image authentication surviving JPEG compression," *Digit. Signal Process.*, vol. 20, no. 6, pp. 1597–1611, 2010.
[19] C.-C. Lai and C.-C. Tsai, "Digital image watermarking using discrete wavelet transform and singular value decomposition," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 11, pp. 3060–3063, Nov. 2010.
[20] M. Urvoy, D. Goudia, and F. Autrusseau, "Perceptual DFT watermarking with improved detection and robustness to geometrical distortions," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 7, pp. 1108–1119, Jul. 2014.
[21] M. Cedillo-Hernández, F. García-Ugalde, M. Nakano-Miyatake, and H. M. Pérez-Meana, "Robust hybrid color image watermarking method based on DFT domain and 2D histogram modification," *Signal, Image Video Process.*, vol. 8, no. 1, pp. 49–63, 2014.
[22] X. Niu, Z. Lu, and S. Sun, "Digital watermarking of still images with gray-level digital watermarks," *IEEE Trans. Consum. Electron.*, vol. 46, no. 1, pp. 137–145, Feb. 2000.
[23] X. Niu, S. Sun, and W. Xiang, "Multiresolution watermarking for video based on gray-level digital watermark," *IEEE Trans. Consum. Electron.*, vol. 46, no. 2, pp. 375–384, May 2000.
[24] R. Shiba, S. Kang, and Y. Aoki, "An image watermarking technique using cellular automata transform," in *Proc. IEEE Region 10 Conf.*, 2004, vol. 1, pp. 303–306.
[25] N. Panahi, M. Amirani, S. Behnia, and P. Ayubi, "A new colour image watermarking scheme using cellular automata transform and schur decomposition," in *Proc. 21st Iran. Conf. Elect. Eng.*, 2013, pp. 1–5.
[26] H. Wu, J. Zhou, and X. Gong, "A novel image watermarking algorithm based on two-dimensional cellular automata transform," in *Proc. 6th IEEE Joint Int. Inf. Technol. Artif. Intell. Conf.*, 2011, vol. 2, pp. 206–210.
[27] O. Lafe, "Data compression and encryption using cellular automata transforms," in *Proc. IEEE Int. Joint Symp. Intell. Syst.*, 1996, pp. 234–241.
[28] O. Lafe, "Cellular automata transforms," in *Proc. Cell. Automata Transforms*, 2000, pp. 23–44.
[29] X.-W. Li and I.-K. Lee, "Robust copyright protection using multiple ownership watermarks," *Opt. Express*, vol. 23, no. 3, pp. 3035–3046, 2015.
[30] X. W. Li, S. J. Cho, and S. T. Kim, "Computational integral imaging-based 3D digital watermarking scheme using cellular automata transform and maximum length cellular automata," *Multidimensional Syst. Signal Process.*, vol. 25, no. 3, pp. 405–424, 2014.
[31] M. Cho and B. Javidi, "Three-dimensional photon counting double-random-phase encryption," *Opt. Lett.*, vol. 38, no. 17, pp. 3198–3201, 2013.
[32] A. Aggoun, "A 3D DCT compression algorithm for omnidirectional integral images," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, vol. 2, pp. II–II.
[33] R. Zaharia, A. Aggoun, and M. McCormick, "Adaptive 3D-DCT compression algorithm for continuous parallax 3d integral imaging," *Signal Process. Image Commun.*, vol. 17, no. 3, pp. 231–242, 2002.
[34] A. Stern and B. Javidi, "3D image sensing and reconstruction with time-division multiplexed computational integral imaging (CII)," in *Proc. IT-Com*, 2003, pp. 131–139.
[35] M. Irani and S. Peleg, "Motion analysis for image enhancement: Resolution, occlusion, and transparency," *J. Vis. Commun. Image Representation*, vol. 4, no. 4, pp. 324–335, 1993.
[36] H. Song, X. He, W. Chen, and Y. Sun, "An improved iterative back-projection algorithm for video super-resolution reconstruction," in *Proc. Symp. Photon. Optoelectron.*, 2010, pp. 1–4.
[37] N. K. Bose and N. A. Ahuja, "Superresolution and noise filtering using moving least squares," *IEEE Trans. Image Process.*, vol. 15, no. 8, pp. 2239–2248, Aug. 2006.
[38] H. Kim, S. Lee, T. Ryu, and J. Yoon, "Superresolution of 3-D computational integral imaging based on moving least square method," *Opt. Express*, vol. 22, no. 23, pp. 28 606–28 622, 2014.
[39] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Robust shift and add approach to superresolution," in *Proc. 48th Annu. Meeting Opt. Sci. Technol.*, 2003, pp. 121–130.
[40] H. Ji and C. Fermüller, "Robust wavelet-based super-resolution reconstruction: Theory and algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 649–660, Apr. 2009.
[41] S. K. Narang and A. Ortega, "Perfect reconstruction two-channel wavelet filter banks for graph structured data," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2786–2799, Jun. 2012.
[42] M. Vetterli and C. Herley, "Wavelets and filter banks: Theory and design," *IEEE Trans. Signal Process.*, vol. 40, no. 9, pp. 2207–2232, Sep. 1992.
[43] C. H. Chang, Z. Ye, and M. Zhang, "Fuzzy-ART based adaptive digital watermarking scheme," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 65–81, Jan. 2005.
[44] M. S. Hsieh and D. C. Tseng, "Image subband coding using fuzzy inference and adaptive quantization," *IEEE Trans. Syst. Man, Cybern. Part B, Cybern.*, vol. 33, no. 3, pp. 509–513, Jun. 2003.
[45] S. M. Youssef, A. A. ElFarag, and N. M. Ghatwary, "Adaptive video watermarking integrating a fuzzy wavelet-based human visual system perceptual model," *Multimedia Tools Appl.*, vol. 73, no. 3, pp. 1545–1573, 2014.
[46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.



Xiao-Wei Li received the Ph.D. degree in information and communications engineering from Pukyong National University, Busan, South Korea, in 2014. From 2014 to 2015, he was a Researcher with Yonsei University, South Korea, working on digital watermarking in the framework of copy protection and image encryption. He is currently an Associate Professor in the School of Electronics and Information Engineering, Sichuan University, Chengdu, China. His research interests include integral imaging, image watermarking, and encryption.



Seok-Tae Kim received the Ph.D. degree from the Department of Communication Engineering, Osaka University, Osaka, Japan, in 1991. He was an Inviting Professor with the University of Washington, USA, in 1999, and an Inviting Professor with Simon Fraser University, Canada, in 2006. Since 1991, he has been a Professor in the Department of Electronics, Computer and Telecommunication Engineering, Pukyong National University, Busan, South Korea. His research interests include image processing, pattern recognition, watermarking, and cellular automata.



Qiong-Hua Wang is currently a Professor of optics in the School of Electronics and Information Engineering, Sichuan University, Chengdu, China. She was a Postdoctoral Research Fellow in the School of Optics/CREOL, University of Central Florida, from 2001 to 2004. She has published more than 200 papers on information displays. Her current research interests include optics and optoelectronics, especially display technologies. She is the Associate Editors of *Optics Express* and *Journal of the Society for Information Display*.

Quantifying Spatial and Angular Resolution of Light-Field 3-D Displays

Péter Tamás Kovács, *Student Member, IEEE*, Robert Bregović, *Member, IEEE*, Atanas Boev, Attila Barsi, and Atanas Gotchev, *Member, IEEE*

***Abstract*—Light-field 3-D displays are expected to play an important role as terminal devices, visualizing 3-D objects apparently floating in the air, or letting viewers see through a window with a scene behind it. Currently, there are neither methods nor practical tools to quantify light-field display's effective resolution or the perceived quality of the presented imagery. Most 3-D displays are simply characterized by the total number of pixels or light rays; however, this number does not properly characterize the distribution of the emitted light rays, or the level of detail that the display can visualize properly. This paper presents methods to measure the spatial (i.e., 2-D equivalent) and angular (i.e., directional) resolution of a given light-field display. The frequency-domain analysis of recorded test patterns gives the spatial resolution limit of the display under test, while angular resolution is determined by the display's ability to pass through patterns of uniform angle-dependent patterns. It also presents a subjective experiment to corroborate the objectively measured spatial resolution.**

***Index Terms*—Displays, distortion, image resolution, spatial resolution, spectral analysis, three-dimensional television.**

I. Introduction

VISUAL information has always been the primary and richest kind of information perceived by humans. Consequently, displays are primary output devices of computers and other electronic devices, including handheld, desktop and large-scale surfaces for representing visual information, some of which have 3D capabilities in recent products. Examples include 3D enabled phones, TVs and computer monitors. With recent advances in display technology, users expect immersiveness and realism when perceiving and interacting with visual information. Thus, displays are aimed at reproducing an increasing subset of visual cues present in reality: vivid colors, high resolution, binocular effect (when two eyes see different images, resulting in stereopsis), and the parallax effect (when an observer can see different perspectives while moving in front of the screen) all contribute to a higher realism of the displayed information [1]. Reproducing binocular effects and motion parallax are unique to autostereoscopic (glasses-free) 3D displays, which are implemented based on various technologies, such as lenticular-lens based multiview [2] or projection-based light-field displays [3]–[6].

A. Motivation

Given different display technologies and models, one needs quantitative measures in order to know what to expect from a given display, what they can deliver in terms of visual realism. The characterization of the throughput of light-field 3D displays is an important yet challenging issue [7].

In case of 2D displays, spatial resolution is one of the major factors affecting visual realism, and is equally important for 3D displays. However, most 3D display manufacturers describe their products using metrics originating from 2D displays, such as the resolution of the underlying display panel, which does not quantify the distribution of these pixels in the spatial/directional domain. Light-field displays do not have a regular pixel structure, as will be described in Section II-A. Therefore measuring the number of features that can be faithfully represented by the display on the screen plane from a single viewing position (i.e., equivalent spatial resolution) is an important metric to judge the visual quality and level of detail presented by the 3D display. The effective resolution away from the screen plane (that is, for content that appears inside the display, or floating in front of the screen) can be calculated from the resolution at the surface using geometrical considerations [3].

Angular resolution is a new metric specific for 3D displays. It does not exist for 2D displays, as those do not exhibit viewing angle dependent behavior. For 3D displays, angular resolution quantifies the quality of the parallax effect. Visible discrete transitions in the motion parallax caused by insufficient angular resolution can hinder viewing experience and may hide important details of the presented information. The smoothness of motion

Manuscript received December 31, 2016; revised May 20, 2017 and June 22, 2017; accepted July 14, 2017. Date of publication August 11, 2017; date of current version October 23, 2017. This work was supported by the PROLIGHT-IAPP Marie Curie Action of the People programme of the European Union's Seventh Framework Programme, REA Grant Agreement 32449. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Yebin Liu. *(Corresponding author: Péter Tamás Kovács.)*

P. T. Kovács is with the Tampere University of Technology, Tampere 33720, Finland, and also with Holografika, Budapest 1192, Hungary (e-mail: peter.t.kovacs@tut.fi).

R. Bregović and A. Gotchev are with the Tampere University of Technology, Tampere 33720, Finland (e-mail: robert.bregovic@tut.fi; atanas.gotchev@tut.fi).

A. Boev is with the Tampere University of Technology, Tampere 33720, Finland, and also with Huawei European Research Centre, Munich 80337, Germany (e-mail: atanas.boev@huawei.com).

A. Barsi is with Holografika, Budapest 1192, Hungary (e-mail: a.barsi@holografika.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2738606

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

parallax largely depends on the number of rays one can see over unit length when moving in front of the display's screen. Even more importantly, angular resolution has a direct effect on the depth range, that is, the range of depth shown with reasonable image quality. By measuring the angular resolution, we are indirectly measuring the maximum depth that can be shown on the display [3].

Spatial and angular resolutions together imply the properties and the necessary processing [8] of the light field content the display should be supplied with.

B. Organization of This Paper

In this paper we propose methodologies for objective measurements to quantify spatial and angular resolutions. Furthermore, we propose and conduct subjective tests to corroborate the results of the spatial resolution measurements. Angular resolution measurement results are validated by technology insights from the display used in the experiments.

The rest of this paper is organized as follows. Section II describes related work in display measurement and briefly introduces light field displays (the 3D display technology this paper is focused on), and the concept for estimating display passband based on the display's geometry. Section III describes the objective measurement methods and how the results are analyzed to calculate spatial and angular resolutions, while Section IV describes the subjective test aimed at corroborating the objective measurement results. Section V presents sample results for a specific light-field display and compares the measurement results with the results of subjective experiment. Section VI concludes the paper.

II. Related Work

A. Light-Field 3D Displays

Projection-based light-field 3D displays as described by Balogh [3] aim to reproduce the light-field of a real or synthetic scene by creating a surface with direction selective light emission. This is achieved by stacking many projection modules in a regular arrangement, so that these modules project light rays onto the screen, typically from behind. All projection engines beam light rays onto the whole screen area and these light rays hit the screen surface in slightly different directions. The special holographic screen onto which the rays are projected lets these rays pass through without changing their direction or mixing the color of the different light rays, as shown in Fig. 1.

The holographic screen applies a limited amount of horizontal diffusion, effectively applying discrete-to-continuous conversion on the discretized sources of light [9].

Using such optical arrangement, it is possible to show different images of the same scene to slightly different directions, so that the two eyes of the viewer can see different perspectives, resulting in stereopsis. Also, yet other images are shown to other directions that are seen when the viewer is moving in front of the screen, resulting in motion parallax. The number of different directions emitted from a single screen position can vary and reproducing hundreds of directions is not uncommon.

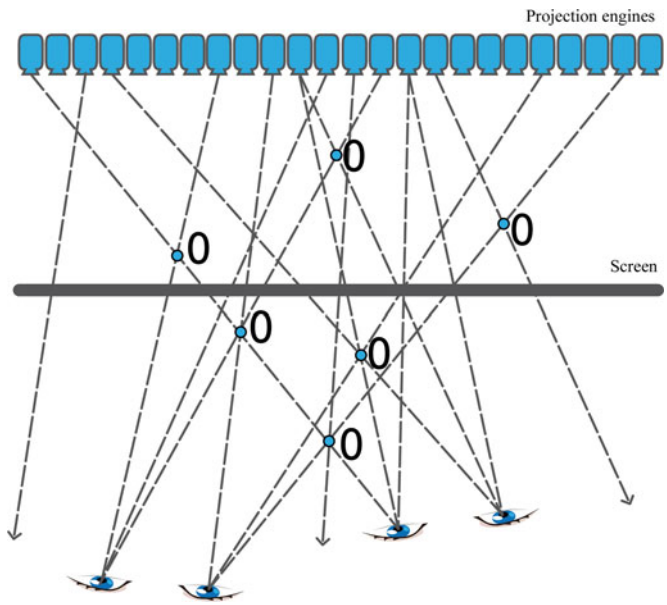


Fig. 1. Principle of operation of projection-based light-field 3D displays. Projection engines project light rays from behind, which may cross each other behind or in front of the holographic screen.

The color of each light ray emitted by the projection modules is determined based on the light field to be represented, and the geometry of the optical arrangement by sampling the desired light field.

The light field is reproduced up to a certain spatial and angular resolution, upper bounded by the total number of light rays emitted by the display.

Projection engines are computer controlled, and thus generating and outputting arbitrary light-fields is possible by software means. The methods described in the next sections are supported by software for generating test light-field patterns, as well as for capturing and analyzing the displayed patterns.

B. Limiting Factors of Light-Field Displays

There are factors that impose an upper limit on the capabilities of projection-based light-field displays, as well as other factors that can affect the final perceived image quality. First, the projection engines generating the light rays, typically containing an imaging component like DLP (Digital Light Processing) or LCoS (Liquid Crystal on Silicon), have finite resolution, such as VGA, XGA, WVGA, HD, or 4k, posing an upper limit on the number of light rays emitted from a single source.

Mounting a large number of projection engines requires a mechanical system, which cannot always ensure pixel-precise matching of light ray hit points on the screen. As a result, positional and rotational misalignments may occur. Also, the optical system of projection engines might have lens distortions, resulting in a non-perfectly rectangular image with equally spaced pixels. While these distortions can be compensated by measuring and pre-distorting the light field, sub-pixel differences still occur, resulting in light rays emitted at fractional positions. Because of this, an observer cannot count pixels on a light-field

display's screen in the way it is possible on a 2D display having a discrete pixel structure.

The finite number of projection engines implies that the number of different directions reproduced is finite, and thus angular resolution has a limit.

While one projection engine can contribute to multiple emitted directions simultaneously using side mirrors which reflect a subset of light rays back to the screen from angles not covered by projection engines, the maximum number of directions cannot be higher than three times the number of projection engines in the system in case of a Horizontal Parallax Only (HPO) system (center, left reflection, right reflection).

C. Display Passband

Spatial and angular resolution are parameters of a more general concept, namely the bandwidth or passband of the display [10]. The passband characterises the display as a light field generator in Fourier domain and as such contains those frequencies, both in the spatial and angular domain (or, equivalently in the ray phase-space domain), which the display is able to reproduce without excessive spatial or inter-perspective aliasing.

As demonstrated in our previous work [9], the passband of a projection-based light field display can be estimated based on the geometrical configuration of the display and applying light field formalism on the display-generated light field. In practice, this requires knowledge about the properties and position of the projection engines and the screen (diffusor). Such approach is particularly useful when designing a display. Based on the desired light field, one can determine the optimal configuration of the display optical elements (which in turn can then be followed by determining optimal camera setup and the necessary pre-filtering) for approximating that desired light field.

In either case (building a display or evaluating the properties of an existing display), the assumption is that the reconstruction support of the diffusor in Fourier domain follows the optimal passband shape. However, in practice, the diffusor cannot have arbitrary shape. It is more-or-less restricted to rectangular shape similar to the ones discussed in [9]. The geometrical analysis considers the ray (spatial-angular) positions and interpret them as sample positions in ray space. As such, it enables estimating the preferable (optimal) reconstruction filter at the diffusor plane.

For illustrative purpose we present here an example. Consider a projection-based display with 22 cpd (cycles per degree) resolution at viewing distance of 3.5 m having 70 views over a field-of-view of 70°.[1]

Following the notations in Fig. 2, this corresponds to spatial-angular sampling at the display plane such that $(x_s, \alpha_s) = (1.40\text{ mm}, 1^\circ)$. Rewriting the relation between the plane with projection units and the screen plane (see [9, eq. (14)]), the parameters describing the sampling grid on the projection plane

[1]The display has been selected such to be close to the real display that will be measured/analyzed later on and for which the real geometrical data is not available.

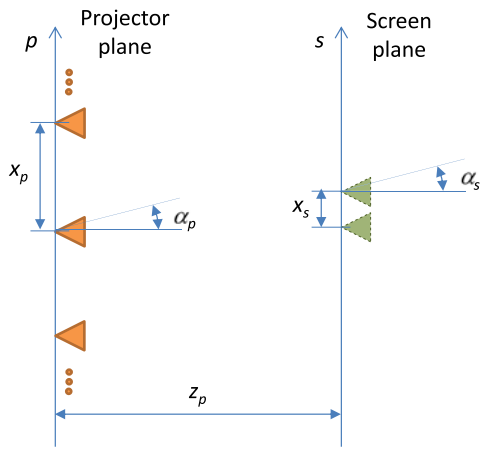


Fig. 2. Projection based light field display setup with notations.

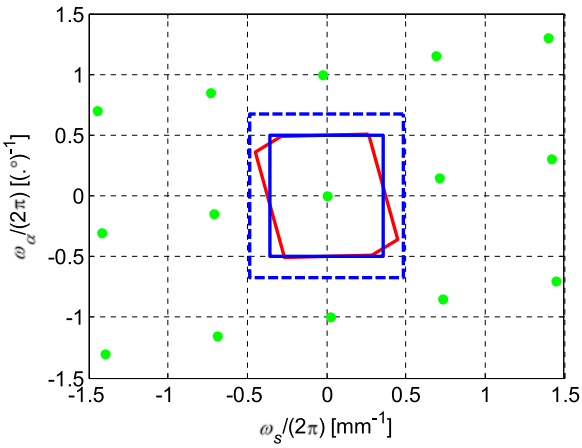


Fig. 3. Frequency domain sampling pattern (green) with estimated display passband (red) and shapes of different diffusors (blue).

are evaluated as

$$\alpha_p = \operatorname{atan}\left(\frac{x_s}{z_p}\right) \text{ and } x_p = \frac{\alpha_s}{\alpha_p} x_s$$

Selecting the distance between screen and projection units as $z_p = 2500$ mm, the display parameters are estimated as $(x_p,\ \alpha_p) = (43.633\text{ mm},\ 0.0321^\circ)$. By performing the analysis as proposed in [9] one obtains sampling pattern in the Fourier domain as shown in Fig. 3. On the same figure, the display passband is shown in red, that in turn would be the optimal shape for the reconstruction filter of the diffusor. In practice the diffusor's Fourier-domain bandwidth is more like the ones of the blue squares with the one shown with solid line being the best candidate for the given sampling pattern. It offers the best overlap with the display bandwidth as determined by the geometry of the ray generators. A diffusor with wider bandwidth would imply worsened spatial selectivity of rays and diffusor with a narrower Fourier-domain support would be unnecessarily restrictive thereby removing finer details.

In summary, there are two issues with the theoretical analysis based on the configuration of optical elements: first, the need to know the correct physical configuration and, second,

to know the properties of the diffusor. From the point of view of a general display user, these are not provided by the manufacturer. Then, the practical solution is to measure the parameters of the display passband in a way that would include the real contribution of all elements building the display. Two measurement approaches aimed at this are introduced in this paper.

D. 3D Display Measurement and Quantification

Most previous work on 3D display measurement have targeted the measurement and quantification of stereoscopic and multiview displays. Boev *et al.* developed a method [11] for modeling multiview displays in the frequency domain and identifying their limits using test patterns of different orientation and frequency. Their work, however only applies to multiview displays and only concerns spatial resolution.

Boher *et al.* developed measurement equipment using Fourier optics to measure the views emitted by desktop multiview displays [12]. While this gives precise results for the targeted class of displays, its applicability for large-scale 3D displays, as well as light-field displays using front projection is rather limited due to the size of the equipment and the way it is used (it would block the light path).

The International Display Measurement Standard (IDMS) published by the Society for Information Display [13] provides guidelines for measuring spatial and angular resolution of 3D displays in general. The recommended angular resolution measurement method relies on showing two-view patterns, which is not applicable for light-field displays, as these do not have discrete views to be controlled. It is also assumed that pixel size is known, which cannot be ensured when measuring an unknown display, which limits the applicability of this method. It further implicitly assumes that the pixels are rectangular, which is not true for several projection technologies (for example diamond shaped pixels in some DLPs).

From the methods presented in the IDMS for spatial resolution measurement for 2D displays, "Resolution from contrast modulation" is probably the closest to our method. This method uses grille lines of discrete sizes and measures the contrast ratio for these patterns. The effective resolution is estimated based on where the contrast ratio is expected to fall below 50%, which is typically located between two discrete measurements. Our method is more precise, because on LF displays patterns with arbitrary size can be shown, thus interpolation is not necessary. The other advantage of our method is that it does not rely on contrast ratio only, but takes into account other sources of noise than decrease in image contrast (regardless of its source).

The authors are not aware of any comprehensive measurement method capable of quantifying spatial and angular resolution of light-field displays or any other 3D display with irregular pixel structure. Our previous work [14] presented an earlier version of spatial resolution measurement of light-field displays and an early method for angular resolution measurement, which used intensity loss for characterizing angular resolution. In this work we expand it by applying frequency analysis for characterizing angular resolution.

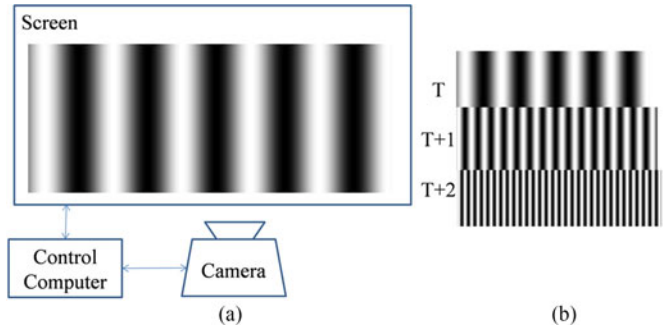


Fig. 4. Spatial resolution measurement overview. (a) A sinusoidal test pattern is rendered on the display under test, while a camera attached to the control computer takes a photo. (b) Subsequent measurement iterations show sinusoidals with increasing frequency.

III. Measurement Methodology

A. Methodology for Spatial Resolution Measurement

Our aim is to devise a methodology for measuring the equivalent spatial resolution of a light-field display using commodity camera and processing tools. The approach is inspired by the methodology presented in [11] for the case of multi-view displays. The measurement is accomplished by (1) visualizing sinusoidal test patterns with different frequency; (2) capturing the visible output by a camera; (3) analyzing the captured images in the frequency domain in order to check if the display reproduces the intended pattern without excessive distortions, and (4) converting frequencies to equivalent spatial resolution.

The test patterns are rendered with custom software that colors each light ray based on the hit point with the display's screen, regardless of its angle.

The rendered test pattern is a black-white sinusoidal that is shown over the whole screen surface (see Fig. 4(a)), initially showing only a few periods on the screen. The attached camera takes a photo of the pattern as shown on the display, and the frequency of the sinusoidal is increased (see Fig. 4(b)). The process is repeated until the frequency of the sinusoid well exceeds the theoretical limit of resolution of the light-field display, which is determined by the resolution of the imaging components. We use frequencies which are up to twice higher than the frequency corresponding to the maximum resolution of the imaging components. The same measurement is repeated for vertical sinusoidals.

The camera recording the test patterns as they are shown is set up on a tripod facing the display; the height of the sensor matches the center of the screen, and the use of manual shooting settings to ensure the captured intensity range matches the brightness of the display, so that the blacks and whites in the displayed patterns are visible on the photos (not saturating the dynamic range of the camera). The camera must be set to a resolution at least $2\times$ times higher than the theoretical maximum resolution of the display (in practice, our algorithm uses four times as many samples to oversample the picture, considering any possible super-resolution effects caused by the multi-projection system). The shutter speed must be slower than the time-multiplexing frequency of the projection components (as projection engines typically employ time multiplexing to emit R, G and B color

components using a single imaging component; this frequency is available in the projection engine's data sheet), and the camera focus is set on the screen plane.

B. Analysis of Spatial Resolution

The analysis of the photos of spatial resolution measurement is based on the observation that the frequency spectrum of the cropped photos, showing only the sinusoidal patterns and the distortions introduced by the display, clearly indicates the limit where the display is incapable of reproducing the test pattern. The test pattern is displayed on the whole screen, photographed, and 1D frequency analysis is performed on a single line of samples in the center of the screen. By increasing the frequency, the amplitude of the dominant frequency decreases, while at the same time distortion and after a certain frequency aliasing are introduced.

The algorithm determines the limit of resolution where the amplitude of distortion exceeds 20% of the amplitude of the dominant frequency. The horizontal spatial resolution then matches with the number of transitions of the test pattern that has been shown before the threshold was exceeded. The 20% threshold is associated with the level of disruptive distortion - this is the level of distortion (originating from aliasing or other nonlinear effects) which makes it impossible for the viewer to precisely identify all features of the presented visual data. The arguments for selecting 20% threshold came from [15], which in turn took insights from previous works, that had reported 10% [16], 15% [17], and 25% [18], respectively.

The same analysis is repeated for the orthogonal direction.

C. Methodology for Angular Resolution Measurement

For performing the desired measurements, one needs to emit different colors or intensities to different directions, and measure the screen from various angles. For a relatively small field of view produced by multiview displays, this may be accomplished by placing a relatively large sensor close to the screen as done in [12], but for the wider field of view typically produced by light-field displays, a moving camera is necessary.

On the display side, one needs to present a light field that has a well recognizable part that looks differently when observed from different angles (see Fig. 5(a)).

In practice, the renderer driving the display uses a GPU shader that receives the parameters of the ray to be colored as input, based on which it calculates the position of the hit point on the screen and forms a rectangle-shaped measured spot. It also calculates the emitted direction, according to which rays are colored either black or white. The screen area that is colored in this way appears black from some viewing angles, and appears white from others, alternating as the observer or camera moves sideways. A camera moving in parallel with the screen plane records the transitions on video, the frames of which are subsequently analyzed. The setup of the camera is similar to that of the spatial resolution test, but the camera is mounted on a motorized rig that allows precise positioning of the camera.

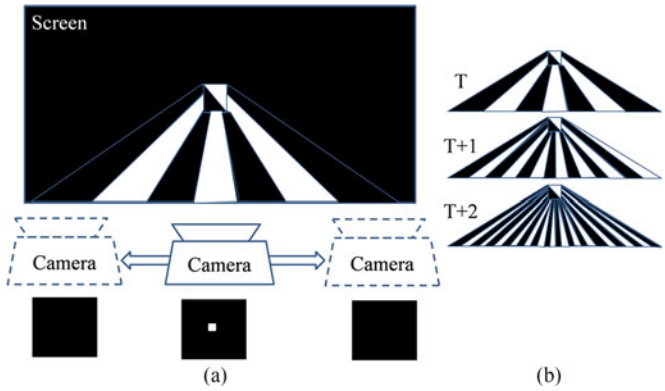


Fig. 5. Angular resolution measurement overview. (a) Test setup with moving camera. The rectangle looks black from some locations and white from other locations. (b) Test patterns of increasing angular frequency.

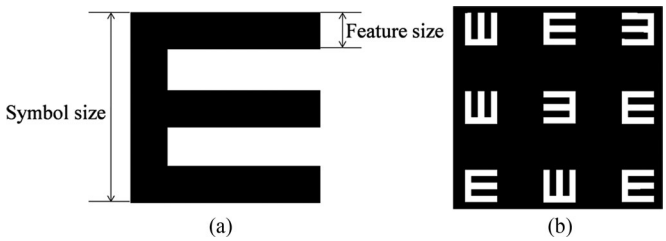


Fig. 6. Subjective spatial resolution test overview. (a) One tumbling "E" symbol. Feature size is 1/5 of the total symbol size. (b) A chart of 9 randomized E symbols arranged in a 3×3 matrix.

In subsequent measurement iterations the angle of black/white features is decreased, thus more transitions are emitted per unit angle, and a new sliding video is recorded (see Fig. 6(b)).

There is a possibility to check the angular resolution relying on information about the engine's order and topology in case of projection-based light-field displays, determining the maximum number of distinct light ray directions that can be emitted. Knowing the order of optical engines, one can force every even projection unit to project white, and odd ones to project black. This presents an intensity profile, while reversing the pattern presents the complement of this profile. Overlaying the two profiles shows peaks at every distinct direction that can be reproduced, and these peaks can be counted (see Fig. 10(b) for an example, where one profile is shown for clarity). As no light can originate from between two discrete light sources inside the display, this limit cannot be exceeded by any test pattern. As our aim is to devise a general methodology for measuring the angular resolution, we use this alternative only to validate the first approach, which in turn requires no direct control of individual projection engines.

D. Analysis of Angular Resolution

The display is considered to be capable of showing the pattern with the given angular resolution if the black/white transitions can be still recognized. In our analysis this is formulated by maximum achievable frequency in the directional domain. As

we increase the angular frequency of the test pattern, we can observe that the transitions are still present; however, after a given frequency, the analysis shows that the dominant frequency is disappearing, or one may even observe decreasing apparent frequency. Frequency analysis is performed on the sequence of photos taken during the measurement. We define the limit of angular resolution where the dominant frequency reaches its upper limit (as seen later in Section V-B).

In case of HPO light-field displays, this measurement is performed in horizontal direction only. However, for 3D displays which have both a horizontal and vertical parallax, the measurement shall be repeated in a vertical direction to characterize the angular resolution in both directions. This is because horizontal and vertical angular resolution might be substantially different.

IV. Subjective Test of Perceived Spatial Resolution

For the case of angular resolution measurement, there is a direct way to verify the correctness of the results of our proposed methods, as discussed in Section III-C. However, there is no such direct method for the case of equivalent spatial resolution measurement. Therefore, we resort to designing and conducting a subjective test to corroborate the proposed objective measurements for that case.

A. Methodology of Subjective Test of Perceived Spatial Resolution

In this test, the perceived spatial resolution of a light-field display is determined by showing visual features with various feature size to participants (see Fig. 6), and asking them to record what they can see on the screen. The visual feature chosen is the tumbling "E" symbol, which is often used by ophthalmologists to measure people's visual acuity [19], thus participants might already be familiar with them.

The methodology, as proposed, consists of two subsequent tests. The first test is done by using symbols printed on paper. It acts as pre-screening aimed at filtering participants with insufficient visual acuity. As such, it ensures that when in later tests people cannot recognize small symbols, this is caused by the peculiarities of the display, and not by the participant's limited visual acuity. In this first test, participants are asked to record the orientation of E's with randomized orientation, printed on paper, arranged in rows of 5, sized from 4 mm down to 0.85 mm, symbol size shrinking with 4/5 in 12 steps (halving every 3 steps), like in real tumbling E charts. From 80 cm viewing distance the 4 mm symbol size corresponds to 0.29 degree symbol size and 0.058 degree ($\sim$3.52 arcmins) feature size (also see Fig. 6(a)). The smallest symbol size (0.8 mm) from the same distance corresponds to 3.69 minutes of arc, and a feature size of 0.73 minutes of arc. Only participants that pass the first test successfully (100% recognition of bigger symbols on paper) proceed to second test.

In the second test, people are presented with the 3D display at the same distance where the paper was shown. The display shows 9 randomized symbols, arranged in a 3×3 matrix (as shown in Fig. 6(b)), and participants copy the orientation of the E's onto a 3×3 matrix on paper. 26 sets are presented in total and the first 4 are considered training sets, while the remaining 22 tests show 11 symbol sizes, both shown twice. Since we are not assessing the visual quality but the ability of the display to discriminate resolution variations, we use similar content for training and the actual experiment in order to allow people to familiarize with the test pattern and the task. The answers given for the training sets are not used during analysis. Symbol sizes are randomized to avoid habituation or anticipation that may affect results obtained by monotonically increasing or decreasing symbol sizes. The time elapsed between the presentation of a new set of symbols and reporting readiness (when all symbols have been recognized and recorded on paper) is measured and logged.

B. Analysis of Subjective Test of Perceived Spatial Resolution

First the results of the paper-based visual acuity tests are calculated, based on which participants with insufficient visual acuity to perform the test are eliminated. After removing the training sets, the randomized pattern sizes are reordered based on the recorded log files, resulting in an ordered list of accuracy and completion time for all participants (except those who failed on the paper based visual acuity test) for 11 symbol sizes of decreasing feature size. The accuracy values for all participants are averaged and the standard deviations and the 95% confidence intervals over recognition data are calculated.

Our hypothesis is that the participants will correctly recognize the symbols up to a given symbol size that corresponds to perceived display resolution and start making significant mistakes (wrong guesses) once the symbols are below the resolution of the display. We aim to identify the subjectively perceived spatial resolution corresponding to the smallest feature size participants are still able to recognize properly. We also check whether the time spent to recognize the symbols increases significantly, which would indicate that the resolution limit has been reached (when participants have difficulties in recognizing symbol orientations).

Although common visual acuity tests determine the limit of visibility where recognition accuracy falls below 50%, as recommended by [19], recent results such as [20] suggest that there may be significant differences between the recognition accuracy of different symbols (optotypes) due to the fact that the unbalanced symbols can be recognized based on their luminance distribution, even if they are heavily blurred or distorted. This suggests that such a subjective test cannot practically determine an absolute limit of visibility. We aim, however, to relate the uncertainty of visibility with the objective measurements.

V. Results and Discussion

We demonstrate the applicability of the proposed measurement approaches on a light-field display. The display under test was a large-scale prototype light-field display with 180 cm screen diagonal. The display is controlled by a rendering cluster with high-end GPUs connected via HDMI connections. The setup for subjective experiments is described in Section V-C.

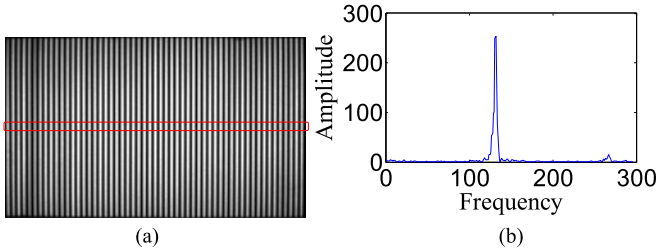


Fig. 7. (a) A photo of the screen showing a sinusoidal test pattern. The center row of the photo is used for frequency analysis. (b) Frequency spectrum of a single measurement showing the sinusoidal, with FFT bins on the horizontal axis.

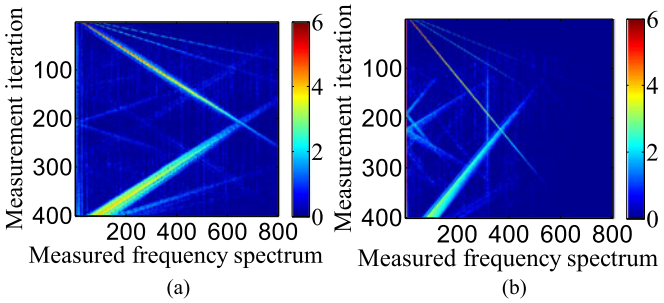


Fig. 8. Frequency spectrums of successive measurements stacked in a matrix. Measurement iteration count increases downwards, while the observed frequency increases rightwards. (a) Spectrums of horizontal resolution measurements. Major sources of distortion are visible as harmonics and constant low-frequency distortion. (b) Spectrums of vertical resolution measurements.

A. *Spatial Resolution Measurement*

During the measurement, 411 photos have been taken for both horizontal and vertical directions and analyzed as described in Section III-B. The analysis procedure has been implemented in a Matlab script.

A photo of the light-field display showing a sinusoidal test pattern is shown in Fig. 7(a). One line of samples is extracted from the center of the screen, on which 1D frequency analysis is performed. The frequency spectrum of a single measurement is shown in Fig. 7(b), with FFT bins shown on the horizontal axis. How these relate to real resolution will be explained later in this section.

Using a series of photos and stacking the resulting frequency spectrums to form a 2D matrix, the frequency response of the display under test with increasing input frequencies is obtained, as shown in Fig. 8. Measurement iterations with increasing input frequency are shown on the vertical axis from top to bottom, while the resulting frequency spectrum is on the horizontal axis. The strong diagonal in the spectrum shows the input frequency being output by the display as the dominant frequency.

There are however other frequency components in the spectrum, the causes of which will be discussed. The decrease of the amplitude of the dominant signal is also visible on the diagonal. After finding the primary and secondary peak for all iterations, the ratio of the amplitude of the primary peak and the amplitude of the secondary peak can be plotted, which shows the level of distortion for each iteration. Such a plot is shown in Fig. 9, where distortion level is plotted in blue against the

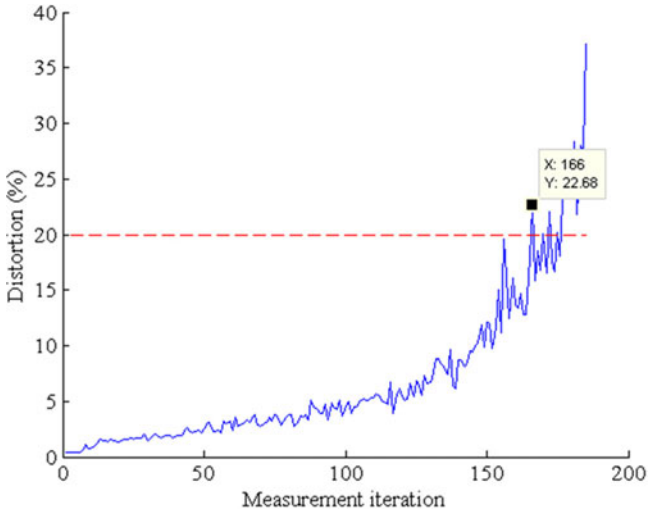


Fig. 9. Level of distortion in subsequent measurement iterations. 20% noise threshold is marked with red dashed line.

20% threshold. From the plot one can see that the first iteration where distortion is above 20% is number 166.

The iteration number can be easily mapped to effective display resolution by plotting the frequency of the primary peak against the measurement iteration.

The resolution is the double of the measured frequency (529 in our case), as one period is considered to be created by the equivalent of two pixels per period in the classical discrete to analog signal conversion (one black and one white). Based on this correspondence between resolution and the frequency the horizontal resolution for the display under test is $2 * (529 - 1) = 1056$ pixels. One is subtracted from the frequency value, as the first bin in the spectrum corresponds to the DC component of the signal.

While determining the effective resolution is a useful result by itself, closely checking the stacked spectrums reveals some sources of distortion. Due to the slightly nonlinear intensity transfer function of the display, sinusoidals appear slightly rectangular, causing harmonics at odd multiples of the dominant frequency, appearing as weaker diagonals, as shown in Fig. 8(a). Also, constant low frequency components on the left-hand side of the spectrum are visible in Fig. 8(a). These are caused by the characteristic of the holographic screen, which results in slightly nonuniform brightness over the screen surface. This manifests itself in a fixed low frequency across all measurements.

Frequency aliasing can also be observed when exceeding the frequency throughput of the display with a large margin. Fig. 8(b) shows the stacked spectrum of the vertical resolution of the measured display, which happens to be lower in the vertical direction. Starting around iteration 170, the mirrored image of the peak appears. Using the same calculation as with the horizontal resolution, the vertical resolution of the display is determined to be 636 pixels.

B. *Angular Resolution Measurement*

Results for the same light-field display as used for the spatial resolution measurement are presented below. During

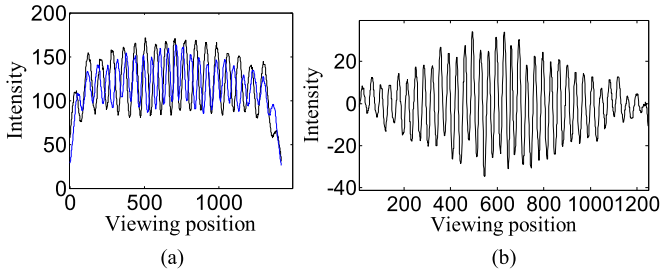


Fig. 10. (a) Sample intensity profiles for two different angular frequencies recorded on the same display. (b) Intensity profiles of forced black-white transitions on a light-field display.

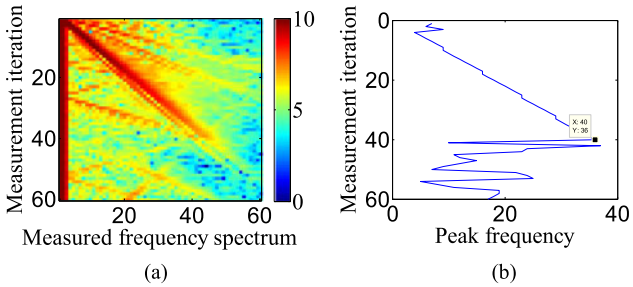


Fig. 11. (a) 1D frequency spectrums of angular resolution test patters stacked in a 2D array. (b) Frequency of the peak in subsequent measurement iterations.

this measurement, 60 videos have been taken with increasing angular frequencies, and analyzed as described in Section III-D. Sample intensity profiles of two different frequencies are shown in Fig. 10(a). Each curve corresponds to the intensity observed by the camera at the center of the captured image as the camera was moving sideways in front of the screen, showing a constant angular resolution test pattern.

By performing frequency analysis on each of the captured intensity profiles and stacking the spectrums below each other (as seen in Fig. 11(a)), the increasing frequency of the test pattern can be clearly observed as a primary peak, with increasing frequency until approx. iteration 45. In subsequent iterations we can see the peak disappear and noise appear instead. Finding the peak in subsequent iterations (as shown in Fig. 11(b)) gives the maximum number of periods one can observe over the field of view of the display. The peak in our case is at 36, which corresponds to 35 full periods (as the first bin in the spectrum corresponds to the DC component of the signal). The number of distinct directions is twice the number of full periods, in our case 70.

Dividing the total viewing angle of the display with this number, the angular resolution can be obtained. This display's total viewing angle is 65.4 degrees according to the measured intensity profile; therefore, the average angular resolution is 0.93 degrees.

Using the direct method for identifying the number of distinct directions that can be reproduced (as described in Section III-C), one can count 35 peaks on Fig. 10(b), showing the intensity profile of the display under test. This alternative approach demonstrates that our method succeeded in finding the maximum number of directions emitted by the display.

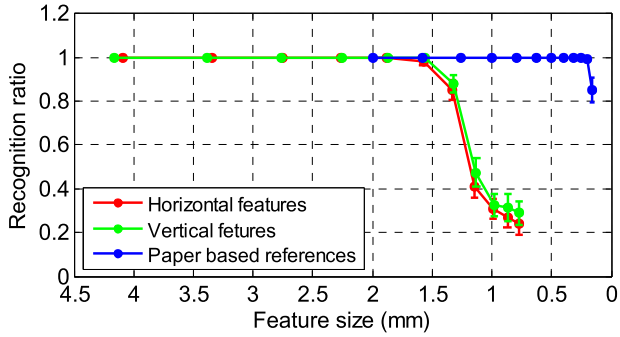


Fig. 12. Recognition accuracy of tumbling E symbols on paper and display, for horizontal and vertical features, plotted against feature size, with 95% confidence intervals.

C. Subjective Test of Perceived Spatial Resolution

Subjective tests have been performed on 53 participants, 43 male and 10 female. The age range of participants was between 22 and 52 years. 21 of them used glasses. All participants claimed to have good eye sight and this was verified using the paper based charts as the first step of the tests. None of the participants were native English speakers (the study being performed in Finland), but represented several nationalities. We selected 50+ participants to participate in the subjective experiment in order to meet the requirements in the ITU Recommendation [21], that specifies that 'at least fifteen participants, non-experts, should be employed'. The distance between the display's screen and the participant's eye was 80 cm during the spatial resolution tests.

In our experiment, 3 out of 53 participants could not reliably record the orientation of the paper based patterns - those were removed from all analysis related to estimating the spatial resolution.

The weighted average of recognition accuracies and corresponding 95% confidence intervals have been calculated for all used symbol sizes for symbols with horizontal and vertical detail orientation (U/D and L/R). Based on symbol size, the feature sizes, and the corresponding resolution values for each symbol size have been calculated. Completion times have been averaged and 95% confidence intervals calculated. The averaged recognition accuracies are shown in Fig. 12. The figures clearly show several effects: the most obvious is that participants recognized the paper based reference almost perfectly (please note that the paper based symbols started smaller, though there is some overlap in symbol sizes). This means the limits visible in recognition accuracy in the case of the display are caused by the display's resolution limit, and not because of the participant's visual acuity limits.

The second noticeable result is that there is a statistically significant drop in recognition around feature size 1.4 mm: there, the confidence intervals do not overlap between measurement steps, and the p-values are very small ($p_{\mathrm{horizontal}} = 3.91e-07$, $p_{\mathrm{vertical}} = 3.54e-07$). This proves our hypothesis that there is a limit on the resolution the display is capable to visualize and it is close to symbol size where the recognition ratios have a major drop.

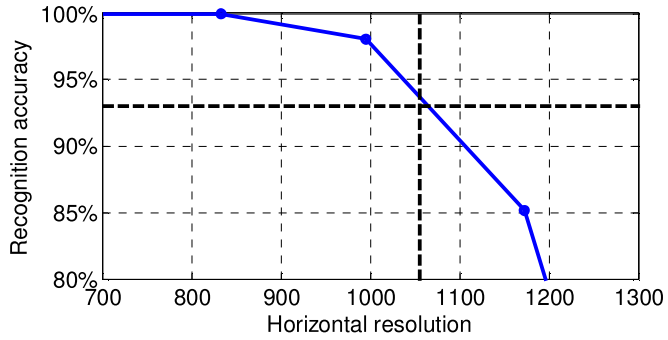


Fig. 13. Zoomed in version of recognition accuracy versus resolution plot for horizontal resolution.

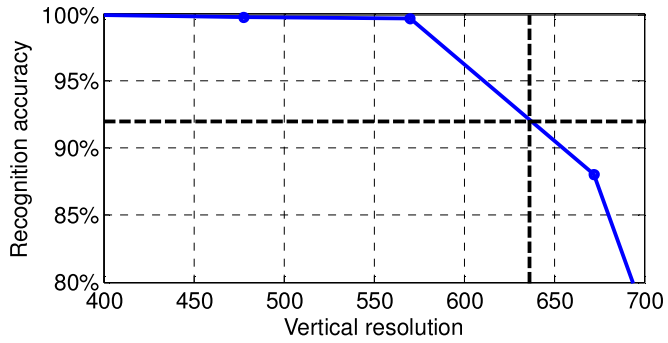


Fig. 14. Zoomed in version of recognition accuracy versus resolution plot for vertical resolution.

From charts in Figs. 13 and 14 it can be read that the measured resolution by the objective method (1056×636) corresponds to 93% and 92% recognition ratios in the horizontal and vertical direction, respectively. This is much higher than the 50% threshold commonly used in visual acuity tests, which can be attributed to the previously mentioned capability of humans recognizing symbols whose features are not entirely visible [20], [22], [23]. Still, the thresholds being so close in the horizontal and vertical directions indicate that the measured resolution values and subjective perception of the visible features are proportional. From a practical perspective, this means that if a content producer provides content complying with the measured spatial resolution, this will ensure good suppression of visually annoying 3D artifacts for the greatest majority of viewers.

Note that the way of carrying out the subjective experiments imposed the use of certain finite set of symbol sizes. Therefore, we used linear interpolation between measured values to estimate the expected recognition accuracy for intermediate resolutions.

A summary of times participants needed to perform each iteration of the test are shown in Fig. 15. It is clearly visible that the average time to recognize and record a group of symbols stays almost constant until feature size $\sim$1.4 mm. Symbols smaller than this limit seem to be increasingly difficult to recognize, indicated by the increased average recognition time as well as the larger confidence intervals. Incidentally, the resolution limit obtained by the objective method is reached at this feature size

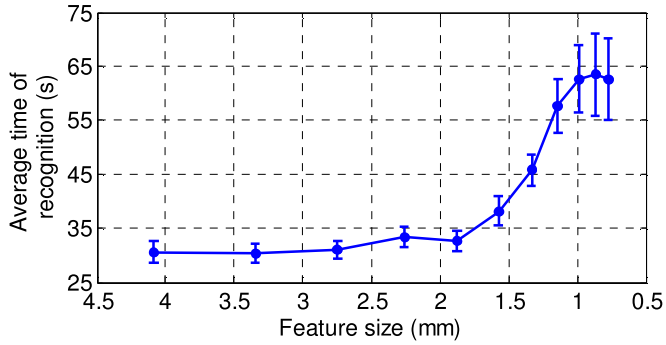


Fig. 15. Average recognition time for a group of symbols with given feature size.

respectively. This shows that a steep increase in recognition time also indicates that the limit of visibility has been reached.

VI. Conclusion

This paper addressed the proper characterization of light-field 3D displays. Two important limiting factors, namely spatial and angular resolution have been discussed, and camera-based objective measurement methodologies have been proposed. A subjective test for corroborating the results of the spatial resolution measurement has also been proposed. These measurement methods and subjective tests have been performed on a light-field display. We have identified limiting factors in light-field displays and how these manifest themselves in the measurements, and how human participants react when these limits are reached. Using the proposed measurement methods, light-field displays can be objectively quantified.

Further work will address any of the advancements over the methods presented here, such as: measuring the perceived resolution on depth planes different from the screen plane; the detection of potential non-uniformities in spatial and angular resolution; or the reduction of measurement time by using smaller number of photos to find the limits.

Acknowledgment

The authors would like to thank the participants of the subjective experiment.

References

[1] M. S. Banks, D. M. Hoffman, J. Kim, and G. Wetzstein, "3D displays," *Annu. Rev. Vis. Sci.*, vol. 2, pp. 397–435, Oct. 2016.
[2] C. van Berkel, D. W. Parker, and A. R. Franklin, "Multiview 3D-LCD," *Proc. SPIE*, vol. 2653, pp. 32–39, Apr. 1996.
[3] T. Balogh, "The HoloVizio system," *Proc. SPIE*, vol. 6055, Jan. 2006, Art. no. 60550U.
[4] G. Wetzstein, D. Lanman, M. Hirsch, and R. Raskar, "Tensor displays: Compressive light field synthesis using multilayer displays with directional backlighting," *ACM Trans. Graph.*, vol. 31, no. 4, Jul. 2012, Art. no. 80.
[5] J.-H. Lee, J. Park, D. Nam, S.-Y. Choi, D.-S. Park, and C.-Y. Kim, "Optimal projector configuration design for 300-mpixel light-field 3D display," *Opt. Exp.*, vol. 21, no. 22, pp. 26820–26835, 2013.
[6] M. Kawakita *et al.*, "Glasses-free 200-view 3D video system for highly realistic communication," presented at the 2013 Digit. Hologr. Three-Dimensional Imag., Kohala Coast, HI, USA, Apr. 2013, Paper DM2A.1.

[7] A. Stern, Y. Yitzhaky, and B. Javidi, "Perceivable light fields: matching the requirements between the human visual system and autostereoscopic 3-D displays," *Proc. IEEE*, vol. 102, no. 10, pp. 1571–1587, Oct. 2014.
[8] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.
[9] R. Bregović, P. T. Kovács, and A. Gotchev, "Optimization of light field display-camera configuration based on display properties in spectral domain," *Opt. Exp.*, vol. 24, no. 3, pp. 3067–3088, Feb. 2016.
[10] M. Zwicker, W. Matusik, F. Durand, H. Pfister, and C. Forlines, "Antialiasing for automultiscopic 3D displays," *ACM Trans. Graph.*, Aug. 2006, Art. no. 107.
[11] A. Boev, R. Bregovic, and A. Gotchev, "Measuring and modeling per-element angular visibility in multi-view displays," *J. Soc. Inf. Display*, vol. 18, no. 9, pp. 686–697, Sep. 2010.
[12] P. Boher, T. Leroux, T. Bignon, and V. Colomb-Patton, "A new way to characterize autostereoscopic 3D displays using Fourier optics instrument," *Proc. SPIE*, vol. 7237, Feb. 2009, Art. no. 72370Z.
[13] *The International Display Measurement Standard v1.03, Society for Information Display*, 2012.
[14] P. T. Kovács, A. Boev, R. Bregović, and A. Gotchev, "Quality measurements of 3D light-field displays," in *Proc. 2014 8th Int. Workshop Video Process. Qual. Metrics Consum. Electron.*, Jan. 2014.
[15] A. Boev, R. Bregović, and A. Gotchev, "Visual-quality evaluation methodology for multiview displays," *Displays*, vol. 33, no. 2, pp. 103–112, Apr. 2012.
[16] L. Wang *et al.*, "Crosstalk evaluation in stereoscopic displays," *Display Technol.*, vol. 7, no. 4, pp. 208–214, Mar. 2011.
[17] P. J. H. Seuntiëns, L. M. J. Meesters, and W. A. IJsselsteijn, "Perceptual attributes of crosstalk in 3D images," *Displays*, vol. 26, nos. 4/5, pp. 177–183, Oct. 2005.
[18] F. Kooi and A. Toet, "Visual comfort of binocular and 3D displays," *Displays*, vol. 25, nos. 2/3, pp. 99–108, Aug. 2004.
[19] International Society for low vision research and rehabilitation, 'Guide for the Evaluation of Visual Impairment', in *Proc. Int. Low Vis. Conf.*, Jul. 1999.
[20] S. P. Heinrich and M. Back, "Resolution acuity versus recognition acuity with landolt-style optotypes," *Graefe's Arch. Clin. Exp. Ophthalmol.*, vol. 251, no. 9. pp. 2235–2241, Sep. 2013.
[21] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R BT.500-13, Jan. 2012.
[22] J. S. Pointer, "Recognition versus Resolution: A Comparison of visual acuity results using two alternative test chart optotype," *J. Optometry*, vol. 1, pp. 65–70, 2008.
[23] L. N. Reich and M Ekabutr, "The effects of optical defocus on the legibility of the tumbling-E and landolt-C," *Optometry Vis. Sci.*, vol. 79, no. 6, pp. 389–393.



Péter Tamás Kovács (S'14) received the M.Sc. degree in computer science from Budapest University of Technology, Budapest, Hungary, in 2004. He is currently working toward the Ph.D. degree in signal processing at Tampere University of Technology, Tampere, Finland.

From 2004 to 2006, he was a Software Engineer with Archi-Data. From 2013 to 2014, he was a Visiting Researcher with Tampere University of Technology. Since 2006, he has been a Software Engineer, a Lead Software Engineer, and then, a CTO with Holografika, Budapest. He is the author or coauthor of four book chapters, four journal papers, and more than 30 conference papers. His research interests include 3-D displays, more specifically light-field displays, and the capture/compression/rendering of light fields.

Mr. Kovács has served as a PC Member for numerous IEEE conferences, a Local Organizing Chair of 3DTV-Con 2014, and is a Contributing Member of the International 3-D Society and the International Committee for Display Metrology, where he contributed to the first IDMS standard. He was the Head of Delegation to MPEG for Hungary.



Robert Bregović (M'97) received the Dipl.Ing. and M.Sc. degrees in electrical engineering from the University of Zagreb, Zagreb, Croatia, in 1994 and 1998, respectively, and the Dr. Sc. (Tech.) degree (with Hons.) in information technology from Tampere University of Technology, Tampere, Finland, in 2003.

From 1994 to 1998, he was in the Department of Electronic Systems and Information Processing, Faculty of Electrical Engineering and Computing, University of Zagreb. Since 1998, he has been with the Laboratory of Signal Processing, Tampere University of Technology. His research interests include the design and implementation of digital filters and filterbanks, multirate signal processing, and topics related to acquisition, processing, modeling, and visualization of 3-D content.



Atanas Boev received the Ph.D. degree from Tampere University of Technology, Tampere, Finland, in 2012. In 2013, he was Visiting Researcher with Holografika KFT, Hungary, involved in development and implementation of light-field rendering algorithms. In 2014, he was a Postdoctoral Researcher with Tampere University of Technology, working on modern signal processing methods for light-field displays. He is currently a Research Engineer with Huawei European Research Centre, Munich, Germany, working on light-field display technology. He is a Researcher with expertize in light-field displays and human stereopsis. He received the best demo award at Tampere Innovation Days in 2010, the best student paper award in Electronic Imaging in 2011, and the Nokia Scolarship award in 2010.



Attila Barsi received the M.Sc. degree in computer science from Budapest University of Technology, Budapest, Hungary, in 2004.

From 2005 to 2006, he was a Software Engineer with DSS Hungary. Since 2006, he has been a Software Engineer, and then, a Lead Software Engineer with Holografika, Budapest. He is the author or coauthor of several conference and journal papers. His research interests include light fields, real-time rendering, ray tracing, global illumination, and GPU computing.



Atanas Gotchev (M'95) received the M.Sc. degrees in radio and television engineering and applied mathematics and the Ph.D. degree in telecommunications from the Technical University of Sofia, Sofia, Bulgaria, in 1990, 1992, and 1996, respectively, and the D.Sc. (Tech.) degree in information technologies from Tampere University of Technology, Tampere, Finland, in 2003.

He is currently a Professor with the Laboratory of Signal Processing and the Director of the Centre for Immersive Visual Technologies, Tampere University of Technology. His recent work concentrates on algorithms for multisensor 3-D scene capture, transform-domain light-field reconstruction, and Fourier analysis of 3-D displays. His research interests include sampling and interpolation theory, and spline and spectral methods with applications to multidimensional signal analysis.

Layered Display with Accommodation Cue Using Scattering Polarizers

Seokil Moon, Chang-Kun Lee, Dukho Lee, Changwon Jang, and Byoungho Lee, *Fellow, IEEE*

***Abstract*—A projection-type layered display inducing the accommodation cue is proposed. To support the accommodation cue, the concept of multiplane displays, especially the principle of additive type compressive light field display, is adopted. To achieve the additive type light field display, two layered scattering polarizers and a single projector which is synchronized with polarization rotator are used. Scattering polarizer diffuses light rays with specific linear polarization state while transmits the other rays with the orthogonal polarization state. Utilizing this characteristic, two layer images can be displayed at different depths presenting multiplane display. By adopting an additional computational process, several viewpoints can be provided to a single eye and support the accommodation cue. We expect that the proposed method can be adopted to virtual reality system. Detailed principle of the proposed system and its feasibility are explained with simulations and experiments.**

***Index Terms*—3D display, layered display, virtual reality.**

I. Introduction

VIRTUAL reality (VR) has received great attention for past few years. Although recent applications for VR are localized to games or entertainment, VR still has a potential to be employed in various fields such as education, medical science and military installations. For instance, computer-based VR simulation technology has a strong possibility to be adopted in neuropsychology since it can integrate the realistic virtual world and human sensitivities without complicated systems [1]. In addition, combat simulators using VR have existed for quite a long time, and through the time the refinement of these simulators has been improved [2]. The main reason for adopting VR technique in such various fields comes from the compact system providing some depth cues. In the past, most of VR systems focused on presenting binocular disparity because it is the most effective depth cue which offers physiological depth perception to the observer [3]. However, providing only binocular disparity makes the observer suffer from a visual fatigue immediately. Creating visually comfortable experiences is crucial to realize the practical VR system and makes it possible to be used in long-term.

The primary reason for the visual fatigue in VR systems is the mismatch between vergence and accommodation. Vergence-accommodation conflict (VAC) appears frequently in stereoscopic three-dimensional (3D) display. Usually, stereoscopic 3D displays do not induce accommodation cue and show interwoven or fused stereo images using a single screen. In this situation, observer's each eye tends to focus on the screen to see clear images, where the observer's brain attempts to make two eyes converge on the position where the 3D image is intended to float. As the conflict between accommodation cue and convergence cue arises, the observer gets visual fatigues [4].

To mitigate VAC problem, accommodation cue should be offered. One simple way to support the cue is providing multiple viewpoints into a single eye. Many research groups have established this condition by adopting multi-projection systems. Lee *et al*. demonstrated a 100-inch multi-projection system with enormous number of projectors and a large-size screen. The viewing angle of the system is 40 degrees and it generates compact view images with excessively high resolution [5]. Takaki *et al.* employed an array of projectors to generate compact viewpoints within the gap of pupil size [6]. However, applying multi-projection systems to VR systems for mitigating VAC problem has some limits. Since each projector can only generate a single viewpoint, a number of calibrated projectors are necessary to provide sufficient number of viewpoints. It would reduce the compactness of the systems and increase their weight and cost. Lee *et al.* reduced the number of projectors in multi-projection system by using uniaxial crystal. According to the polarization state of a single projection unit, double refraction occurs in the uniaxial crystal and it separates the optical path into two different paths [7]. Although the system reduces the number of projectors into half, it is too complicated and expensive to be adopted for VR systems.

Another way to induce accommodation cue is realization of multi-plane displays. Depth-fused displays (DFD) are well-known multi-plane displays which provide accommodation cue by exploiting physiology characteristics of the human eye. When the observer perceives depth-fused 3D images, the focusing position is adjusted to give the maximum contrast of the retinal image. In that case, the perceived depth of fused pixel is determined according to the luminance ratio between superimposed pixels [8], [9]. DFDs are applicable to VR systems as a solution of VAC problem since they can represent small form factor.

Manuscript received December 1, 2016; revised April 30, 2017 and June 15, 2017; accepted July 20, 2017. Date of publication August 11, 2017; date of current version October 23, 2017. This work was supported by "The Cross-Ministry Giga KOREA Project" of, ICT and Future Planning, Korea (GK16D0200, Development of Super Multi-View Display Providing Real-Time Interaction). The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Yebin Liu. *(Corresponding author: Byoungho Lee.)*

The authors are with the School of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea (e-mail: kyo517@snu.ac.kr; lastmeal0929@gmail.com; ldh991@snu.ac.kr; wkdckddnjs@gmail.com; byoungho@snu.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2738614

1932-4553 © 2017 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

DFDs, however, present only a single viewpoint which means the observer cannot perceive the view images outside of the fixed position. This feature would bring discomfort to users because during the long-term use of VR devices, pupil swim is inevitable. Furthermore, as the position of viewpoints are fixed, it is difficult to fit multiple users with various inter-pupillary gaps. Although the additional techniques such as pupil tracking can mitigate these problems, they make the system complicated and expensive [10]. To provide a number of view images with multi-plane displays, systems called compressive light field displays have been proposed [11]–[15]. Light fields are collections of rays indicating a 3D scene from different perspectives. To provide high-resolution light fields, displays usually require an excessively high resolution. On the other hand, compressive light field displays are able to offer high-resolution scenes with computational-processing algorithms such as nonnegative matrix factorization. In particular, using additional optical components, it is possible to build VR system which generates compact viewpoints inducing accommodation cue [16]. However, stacking LCD panels inevitably brings resolution loss due to diffraction from pixelated structure in the LCD panel. Since a degree of diffraction intensifies as the size of pixelated structure decreases, it hinders to stack high-resolution display panels. To overcome this limitation, our group recently published the concept of additive light field display. It handles the suitable algorithm for display systems with stacked panels with non-pixelated structure [17]. By combining this algorithm with characteristics of anisotropic material, it is possible to generate another type of multi-plane displays [18]. Although it can achieve dense viewpoints with small form factor, it is too expensive to construct the system.

In this paper, a new type of compressive light field display which supports accommodation cue is proposed by employing stacked scattering polarizers. The system consists of two sheets of scattering polarizers and a single projector which is synchronized with polarization rotator. Scattering polarizer diffuses light rays with specific linear polarization state while transmits the light rays with the orthogonal polarization state. The projector generates the first layer image and the second layer image in sequence and the synchronized projector sets these two images to have orthogonal polarization states. As two different images are diffused in different depth positions, it is possible to adopt the concept of compressive light field display. Different from conventional compressive light field displays, the proposed system has a potential to overcome the diffraction issue because scattering polarizer does not have pixelated structure. Furthermore, it is easy to construct the system with low cost since it employs a single projector. By virtue of these characteristics, the proposed method has a potential to be adopted in VR system.

II. Light Field Synthesis

A. Scattering Polarizer

The system adopts scattering polarizer as an optical diffuser to realize the multi-plane display. As presented in Fig. 1, it diffuses light rays having specific linear polarization state while transmits light rays with orthogonal linear polarization state.

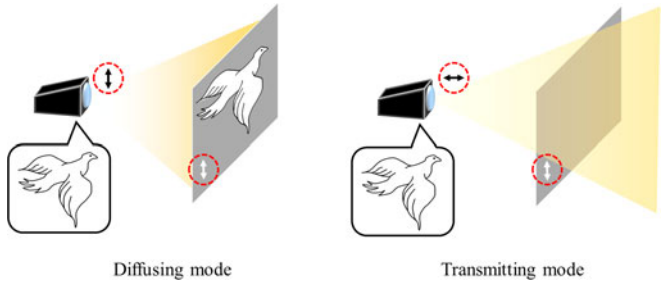


Fig. 1. Characteristic of scattering polarizer. It diffuses light rays with specific linear polarization state while transmits light rays having orthogonal linear polarization state.

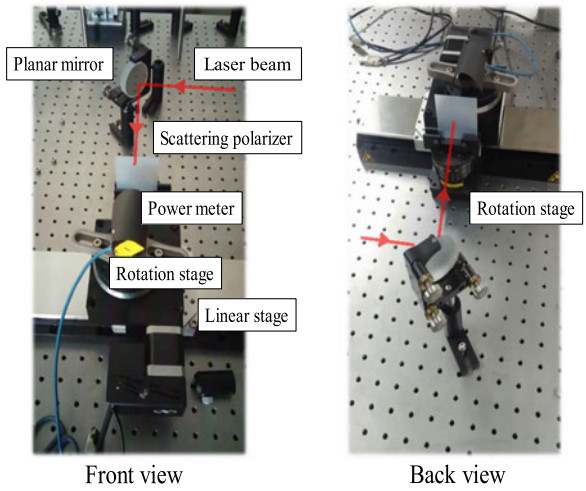


Fig. 2. Experimental setup to analyze the diffusing characteristic of scattering polarizer. Power meter is placed on the rotation stage which is combined with linear stage. As the linearly polarized laser beam is projected onto the scattering polarizer, it diffuses to all directions. The power measures the intensity of diffusing light rays.

When the scattering polarizer acts as a diffuser, it spreads light rays as conventional light shaping diffusers. To confirm the diffusing characteristic of the scattering polarizer, brief experiments are implemented as Fig. 2. Power meter is placed on the rotation stage which is combined with the linear stage. In front of the power meter, a small piece of scattering polarizer is placed. Linearly polarized laser beam is projected onto the scattering polarizer and diffuses as it passes through the scattering polarizer. The power meter moves and rotates on the stages to measure the intensity of the light ray propagating with specific angle.

Fig. 3 represents the diffusion characteristics of scattering polarizer and conventional 10-degree beam shaping diffuser. As the diverging angle of the light ray increases, its intensity decreases with general diffusion characteristic. We assumed that the scattering characteristic curves for the laser beams with orthogonal linear polarization states are identical.

It is well known that intensity of scattered light rays follows Lambertian distribution as the diffusing angle increases [19]. Thus, it is reasonable to use scattering polarizer as a diffuser in the proposed system. The transmitting characteristic of scattering polarizer is also analyzed. In ideal case, if the polarization state of the projected laser beam is orthogonal to the polarization

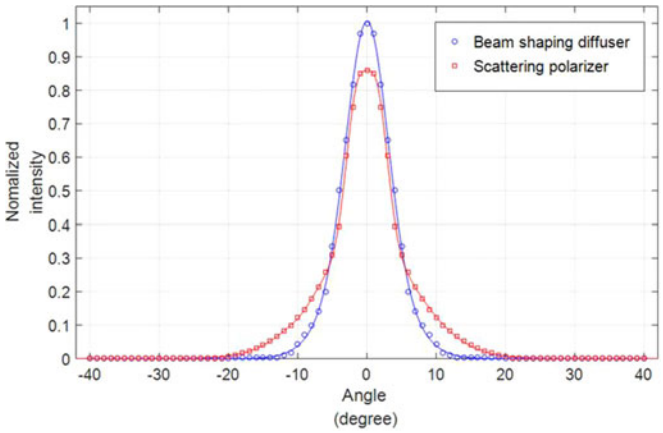


Fig. 3. Diffusion characteristic of scattering polarizer and conventional 10-degree beam shaping diffuser. As shown in the graph, the intensity of scattered light rays from scattering polarizer follows the general diffusion curve. Also the shape of the diffusion curve of scattering polarizer is similar to that of the conventional beam shaping diffuser.

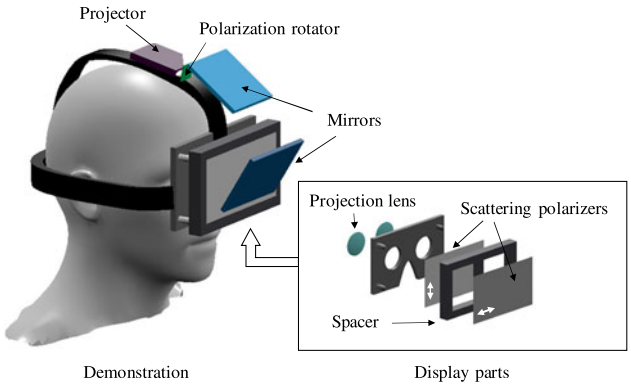


Fig. 4. Schematic of VR system using layered scattering polarizers. Two images are sequentially projected and the synchronized polarization rotator sets these images to have orthogonal linear polarization states. Each of image is diffused by two stacked scattering polarizers respectively.

state of scattering polarizer with diffusing mode, 100% of laser beam should be transmitted. However, according to the characteristic of the scattering polarizer itself, some portion of light which ought to be transmitted is diffused. This characteristic is analyzed by the power meter. 49% of beam is diffused and remaining 51% beam is transmitted through the scattering polarizer. This characteristic induces image contrast degradation when two scattering polarizers are stacked. Detailed explanation about the contrast degradation will be covered in Section III-C.

B. Schematic of the Proposed System

According to the characteristic of scattering polarizer, it is possible to realize a multi-plane display. Fig. 4 represents the schematic of the proposed VR system.

A single projector which is synchronized with polarization rotator generates the front layer image and the rear layer image in sequence. Mirror group is placed in front of the projector to fold the optical path and make light rays to be projected onto layered scattering polarizers. When the front layer image is projected,

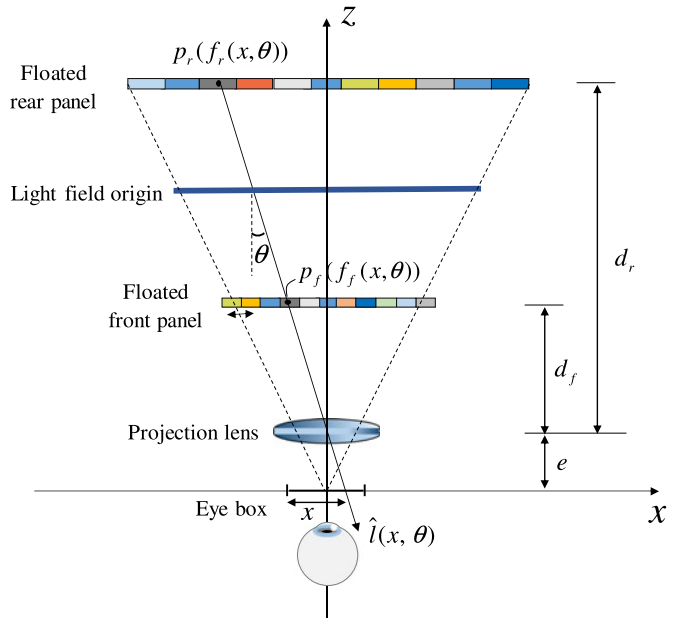


Fig. 5. Equivalent model of the proposed VR system for monocular case. The light field from the origin passes through floated panels. The intensity of the light field is determined by the summation value of two pixel intensities.

the polarization rotator is set to modulate the light into vertically linear polarization state. In the next sequence, the second layer image is projected and the polarization rotator modulates the light into horizontally linear polarization state. The rear scattering polarizer transmits the vertically polarized component and diffuses the image with linear polarization state in horizontal direction. The transmitted image from the rear scattering polarizer is diffused by the front scattering polarizer. Two diffused images from each of scattering polarizers are synthesized and projected onto observer's eyes.

C. Light Field Optimization

If the distance between the projection lens and stacked scattering polarizers is smaller than the focal length of lens, two virtual images of each scattering polarizers are generated for monocular eye. Fig. 5 shows the equivalent model of the proposed VR system.

Derivation of equations for light field optimization is well investigated in previous research [11]–[15]. In the proposed system, two parameters x and θ are adopted to express the arbitrary light field. x is the spatial coordinate on the observing plane and θ is the diverging angle of the light field. The intensity of the pixel in the floated rear panel is expressed as $p_r(f_r(x,\theta))$ where $f_r : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is a function which maps each light ray (x, θ) on the eye box to the floated rear panel. In similar way, for p_f, $f_f : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ maps the ray (x, θ) to the floated front panel [15]. d_f is the distance between eye box and floated front panel, and d_r is the distance between eye box and floated rear panel. e is the eye relief distance. Using simple geometrical relationship, we can derive each mapping function as below:

$$f_f(x,\ \theta) = \frac{d_f + e}{\tan\left(\frac{\pi}{2} - \theta\right)} + x\,, \tag{1}$$

$$f_f(x,\ \theta) = \frac{d_f + e}{\tan\left(\frac{\pi}{2} - \theta\right)} + x\,, \tag{2}$$

where d_f and d_r can be easily derived by lens equation. The emitted light field $\hat{l}(x, v)$ is expressed as follows:

$$\widehat{l}(x,\ \theta) = p_f(f_f(x,\ \theta)) + p_r(f_r(x,\ \theta))\,. \tag{3}$$

To solve (3) for every light field in the system, discretization is progressed as follows:

$$\widehat{\mathbf{L}} = (\mathbf{F}_f\,\mathbf{p}_f) + (\mathbf{F}_r\,\mathbf{p}_r)\ = \mathbf{F}\,\mathbf{p}\,, \tag{4}$$

where $\mathbf{F}_f$ and $\mathbf{F}_r$ are projection matrices that permute the rows of the discrete floated panels $\mathbf{p}_f$ and $\mathbf{p}_r$ according to the mapping functions f_f and f_r, respectively [15]. If a target light field $\mathbf{L}$ is given, an optimization problem can be expressed in a least-squared error sense where $\mathbf{p}_f$ and $\mathbf{p}_r$ provide positive constraints:

$$\min \|\,\mathbf{L} - \ \mathbf{F}\,\mathbf{p}\|^{\,2}\ . \tag{5}$$

Simultaneous algebraic reconstruction technique (SART) is adopted to solve (5) [13], [20].

D. Simulation: Light Field Generation

To investigate the feasibility of the proposed VR system, simulations for several target light fields are proceeded. The spatial resolution of target light fields is 740 by 740. The focal length of projection lens is 65 mm. Distance from the lens to front scattering polarizer is set to 56.5 mm and the physical gap between two scattering polarizers is set to 4 mm. Eye relief distance is neglected. In this situation, two virtual panels are placed at 432 mm (2.315 D) and 874 mm (1.144 D) by projection lens. Fig. 6 represents the simulation results. As shown in Fig. 6(a), two image sources (four different types of sports balls and figure of three men) are rendered from Blender [21]. In the first target image, tennis ball and basketball are about 432 mm and 874 mm away from the virtual camera, respectively. Soccer ball and volley ball are located in the middle point between other two balls. For the second target image, red painted man is standing 432 mm away from the virtual camera, while the streetlamp is located 874 mm away from the virtual camera. As shown in Fig. 6(b), 9 view images having slightly different parallax are projected onto each eye. The virtual camera takes each view image with 2 mm gap. Distance between center view images for left and right eyes is 62.5 mm which is similar to inter-pupillary distance. Fig. 6(c) represents generated layer images for both target light fields. The observer perceives light fields synthesized by two layer images.

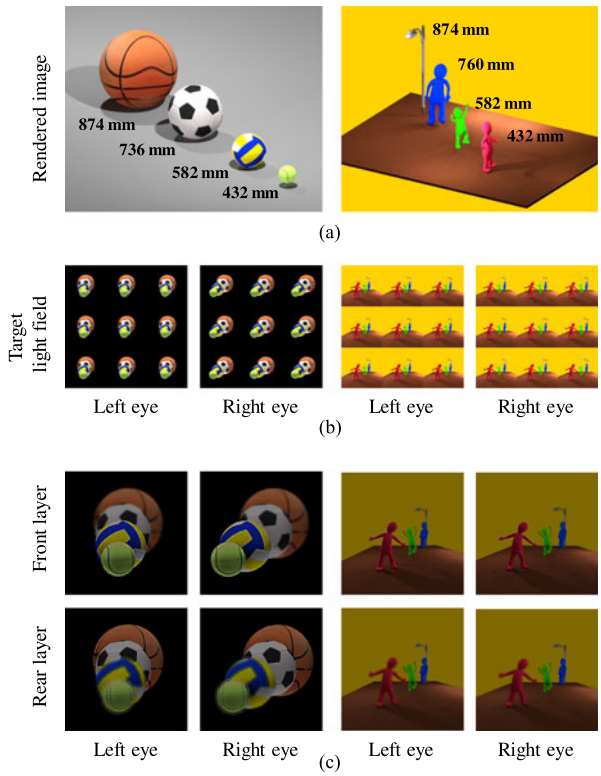


Fig. 6. Configuration of (a) two types of rendered images, (b) target light fields projected onto observer's eyes and (c) rear layer images and front layer images for VR system.

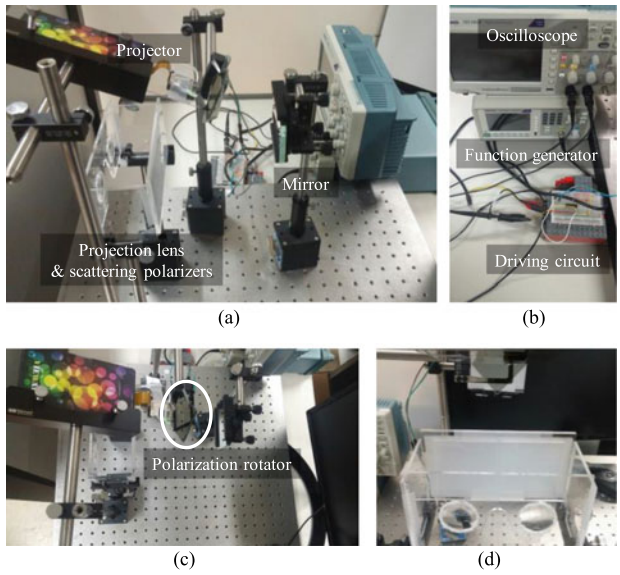


Fig. 7. Experimental setup of the VR system using scattering polarizers. (a) A single projector, polarization rotator, mirror, and layered scattering polarizers are adopted. (b) Voltage signal is applied through the driving circuit and function generator. (c) Top view and (d) front view of the proposed system.

III. Experiments and Analysis

A. Experimental Setup

For investigating the feasibility of the VR system using scattering polarizers, experiments for two target light fields are proceeded. Experimental setup is presented in Fig. 7. As shown in Fig. 6(a), a single projector (Viewway VPL-201, Oriental Electronics, Korea) is placed to provide two layer images in sequence. Behind the projector, the synchronized polarization rotator (2x2_X-FPM[L], LC-Tec displays, Sweden) is placed. It rotates inner liquid crystal (LC) molecules to make each of layer images has appropriate polarization state. Light fields are reflected by the mirror after they pass through the polarization

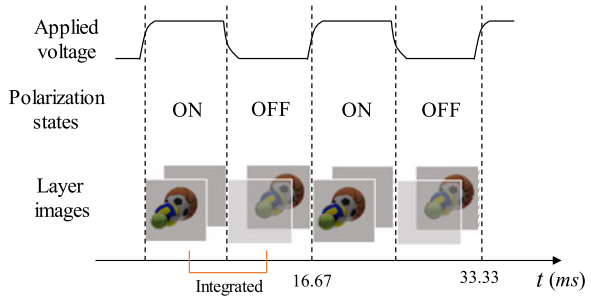


Fig. 8. Configuration of real-time synchronization. The LC molecules inside of the polarization rotator are controlled by applied voltage swing. During the LC response, two layer images are projected in sequence appropriately.

TABLE I
EXPERIMENTAL CONDITIONS

Specifications	Values
Distance between projection lens and front scattering polarizer	56.5 mm
Gap between two scattering polarizers	4 mm
Focal length of the projection lens	65 mm
Distance between projection lens and floated front panel	432 mm
Distance between projection lens and floated rear panel	873 mm
Projection distance (Projector → Mirror → Scattering polarizers)	234 mm
Pixel pitch at scattering polarizer	74.3 um
Resolution of layer images	740 by 740

rotator and diffused by layered scattering polarizers (Teijin Dupont Films, Japan). Fig. 7(b) represents the driving circuit with function generator to apply a voltage swing to the polarization rotator. Synchronization between the projector and the polarization rotator is established by comparing two signals using oscilloscope. The ideal speed of the projector is 60 Hz, but the actual frequency of the signal from the projector is 59.976 Hz. This signal is copied and applied to the rotator using function generator. The LC response time of the rotator is less than 30 μs which is negligible [21]. Fig. 7(c) and (d) are top view and front view of the system, respectively. Fig. 8 is the detailed configuration of real-time synchronization for each of optical components. The LC response inside of the polarization rotator is controlled by applied voltage swing. During the LC response, two layer images are projected in sequence. After the synchronization, two layer images are displayed at desired scattering polarizers. The detailed experimental conditions are written in Table I.

B. Image Distortion Compensation

In the proposed system, two types of distortion occur during the projection process as shown in Fig. 9(a). First, keystone distortion arises because light rays from the projector perform oblique incidence after they are reflected from the mirror. The second distortion is pincushion distortion which occurs by the projection lens. H_1 and H_2 represent the transformation matrices for keystone distortion and pincushion distortion, respectively. For the input image I_0, distorted output image I_{dis} is expressed as below:

$$I_{dis} = H_2\, H_1\, I_0\,. \tag{6}$$

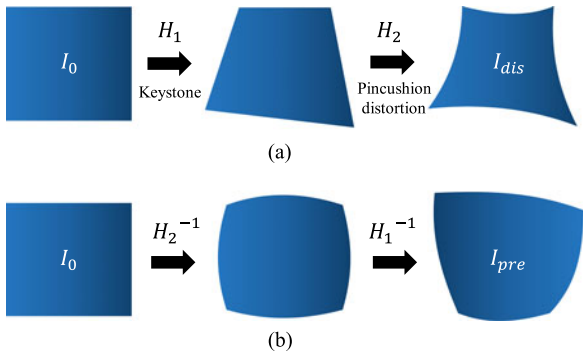


Fig. 9. Image distortion in the proposed system. (a) Two different types of distortion in the system and (b) pre-distortion process.

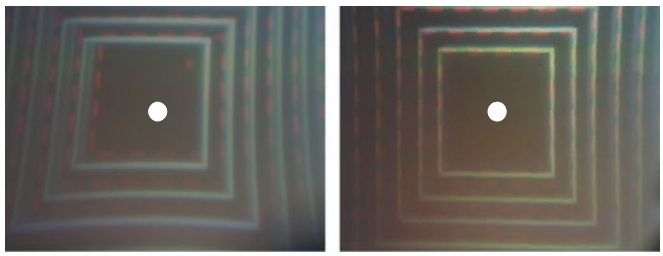


Fig. 10. Image compensation for keystone distortion and pincushion distortion. Observed sample images (a) before the compensation and (b) after the compensation. White lines and red dashed lines represent the front layer image and the rear layer image, respectively.

Fig. 9(b) shows the image compensation process using pre-distortion method. Pre-distorted image I_{pre} is determined as follows:

$$I_{pre} = H_1^{-1}\, H_2^{-1}\, I_0\,. \tag{7}$$

To compensate the keystone distortion, planar homography method is adapted [22]. By comparing 2D coordinate values of four corner points between the original image and the transformed image, it is possible to derive the transformation matrix H_1 as a 3 by 3 matrix.

There are various types of compensation methods for pincushion distortion. In this paper, degree of pincushion distortion is estimated using a checkerboard image. Ma *et al.* suggested several radial distortion models which are radially symmetric around the image center [23]. If a point in $(x,\, y)$ space with radial distance r transforms into another point in $(u,\, v)$ space, the distorted radial distance $\hat{r}$ is expressed as follows:

$$\widehat{r} = \frac{1 + k_1 r + k_2 r^2}{1 + k_3 r + k_4 r^2 + k_5 r^3}\, r, \tag{8}$$

where $k_i\ (i = 1, 2, \ldots, 5)$ are constants [23]. By detecting lattice points at the checkerboard, it is possible to estimate optimal values of k_i and ultimately derive H_2.

Overall pre-distortion is applied to sample images to compensate both keystone distortion and pincushion distortion as shown in Fig. 10(a). For each of layer images, two independent

pre-distortions are applied. Two layer images are well-organized after the pre-distortion as presented in Fig. 10(b).

White lines represent the front layer image and red dashed lines represent the rear layer image. White circles in both Fig. 10(a) and (b) present the center of scattering polarizers. After the compensation, two different lines are perfectly overlapped based on the center circle. As shown in Fig. 10, when observing the layer images contrast degradation occurs. When the rear scattering polarizer is aligned to transmit the front layer image, only 51% of the light is transmitted. Remaining 49% of the light is diffused in bidirectional way. Hence, 24.5% of the light indicating the front layer image is diffused at the rear scattering polarizer and propagates toward the front layer. When it reaches the front scattering polarizer, it is diffused again. Twice diffused light does not deliver any information but degrades the contrast of layer image. In the next sequence, when the rear scattering polarizer diffuses the rear layer image, 50% of diffused light propagates toward the front layer (since the light is diffused in bidirectional way). Then, 49% out of diffused light is diffused again in the front scattering polarizer. In other words, 24.5% of rear layer image is also diffused twice and image degradation occurs. This image degradation due to the characteristic of scattering polarizer is inevitable, but it can be relieved by adopting scattering polarizers with definite characteristic. For instance, if the transmittance of scattering polarizer for transmitting mode is 90%, only 5% of the light degrades the image quality. The practical contrast ratio of the system is measured. ANSI contrast method is adopted to calculate the contrast of the system. The 4 by 4 checkerboard image is projected and average luminance ratio between black parts and white parts is achieved. The measured contrast ratio of the proposed system is 10.85:1.

C. Experimental Results

Fig. 11 represents the results of the experiment. Binocular images for two types of target light fields are used to detect focal sweep. The spatial resolution of target light fields is 740 by 740. Nine viewpoints with 2 mm gap are projected to a single eye box. We used CCD camera and F#1.4 lens with focal length of 25 mm to capture the light fields and clearly show the possibility that the system can support the accommodation cue. The first target light fields consist of four sports balls. Same as the simulation condition, tennis ball and basketball are about 432 mm and 874 mm away from the camera, respectively. Other two balls are located between them. When the camera lens focuses on the tennis ball which is placed foremost among the balls, basketball is blurred. On the contrary, if the camera lens focuses on the basketball, tennis ball and volley ball are relatively blurred. In the second light fields, three men are standing in different locations and a streetlamp stands behind them. The foremost man (red colored) is standing 432 mm away from the camera, and the streetlamp is located 874 mm away from the camera. If the foremost man is focused, the streetlamp is blurred and foremost man is blurred as the camera focuses on the streetlamp.

Fig. 12 demonstrates the experimental results when focusing on 3D images on the front, middle and rear positions. The upper end of the volley ball is located 582 mm away from the camera. When the camera focuses on the middle part among four balls, the upper end of the volley ball is clearly detected compared to other two cases.

D. Diffraction Analysis and Resolution Limits of the System

The primary aim of the proposed VR system is the support of the accommodation cue. When it comes to realizing an accurate and clear accommodation cue, diffraction effect in pixelated structures is a crucial hindrance. For this reason, stacking multi LCD panels to achieve the VR system providing accommodation cue is limited in practice. Furthermore, since the degree of diffraction is in inverse proportion to the size of pixel structure, it is difficult to construct high-resolution VR devices using stacked LCD panels. On the contrary, the proposed VR system is relatively free from the diffraction effect since layered scattering polarizers diffuse light rays without pixel structures. Detailed analysis about the diffraction effect is presented in the following.

As a rectangular aperture with width w and height h is illuminated, the intensity pattern that appears along $x-y$ plane caused by diffraction is expressed as multiplication of two sinc functions. If the distance z between the aperture and intensity pattern is sufficiently large compared to x and y, intensity pattern can be expressed as (9) [24]:

$$I_i(\theta_x, \theta_y) \approx I_c \operatorname{sinc}^2\left(\frac{w}{\lambda}\theta_x\right) \operatorname{sinc}^2\left(\frac{h}{\lambda}\theta_y\right). \tag{9}$$

Here, angles θ_x and θ_y are $\tan^{-1}(x/z)$ and $\tan^{-1}(y/z)$, respectively. λ is wavelength of the light and I_c is intensity of incident plane wave. As the wave propagates through the floated rear panel, it spreads in a cone shape due to the rectangular aperture of the pixel. Before the diffraction is considered, each ray passes through a single pixel of the floated front panel. However, in the real situation, a set of light rays are involved from a single pixel of the floated front panel due to the diffraction. These light rays would affect neighboring pixels at the floated rear panel. Fig. 13 represents the diffraction pattern induced by the pixel at the front panel for the proposed system parameters. Three lines represent the boundaries of $0.9I_c$, $0.5I_c$, and $0.1I_c$, respectively. White lines present the pixel boundaries of the floated rear panel. In this case, adjacent 8 pixels are affected by the pixel at the front panel with 10% scale. To analyze the diffraction effect in more precise way, the projection matrix is modified by calculating the convolution between the original projection matrix and a scale matrix which considers the intensity of adjacent 8 pixels with 10% scale. Previous peak signal-to-noise ratio (PSNR) value for reconstructed images of four balls is 30.2511. After the projection matrix is modified according to the diffraction effect based on (9), PSNR value becomes 29.7942. This indicates that the diffraction effect involves image degradation and affects image quality.

Furthermore, PSNR values are calculated when there is a diffraction effect and there is not, for various resolutions. PSNR differences between these two cases are calculated as shown in Fig. 14. As represented in the figure, as the resolution increases, image degradation due to the diffraction effect intensifies.

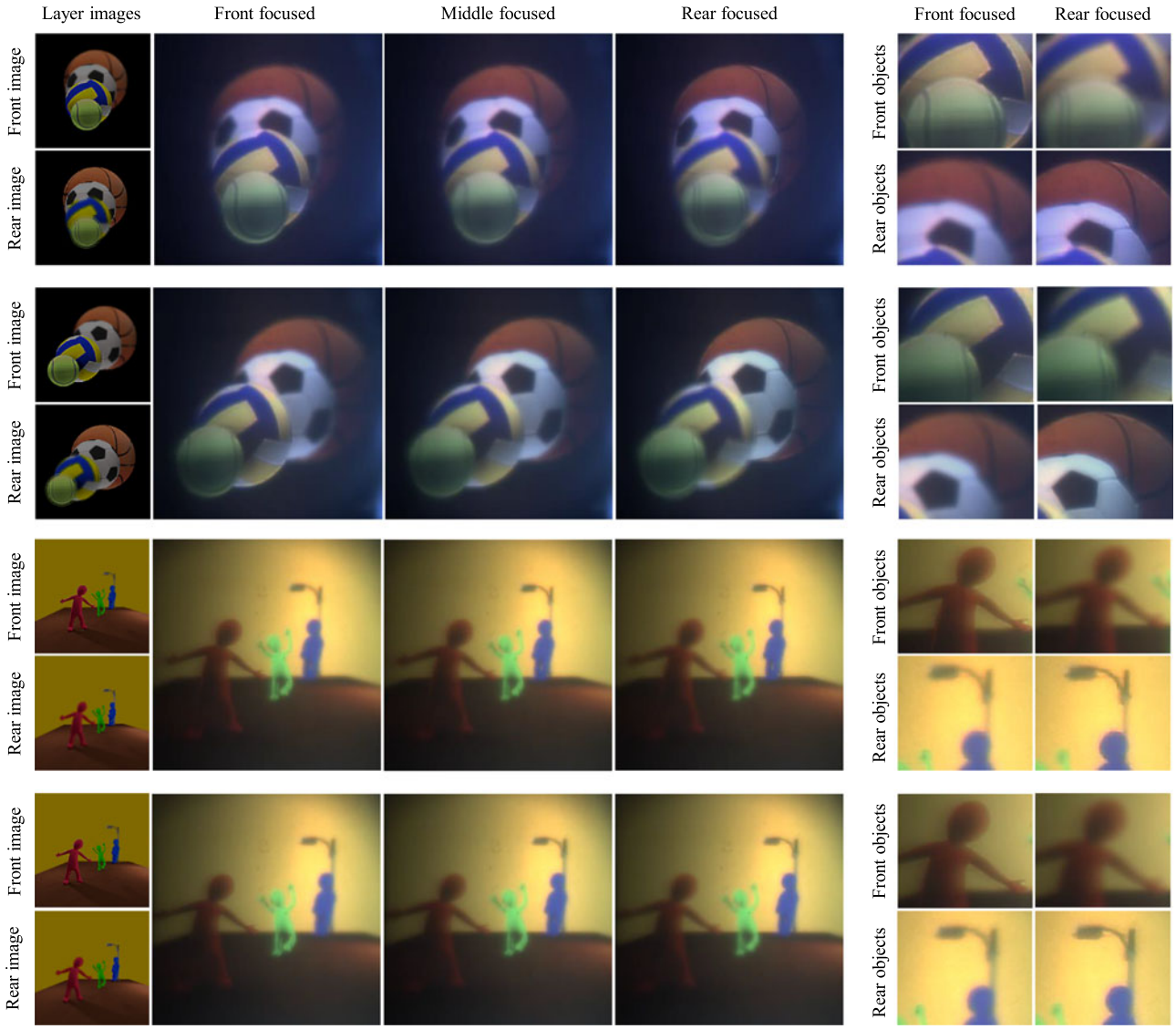


Fig. 11. Experimental results for two different target light fields. In the first target light fields, tennis ball and basketball are located on the floated front and rear panels, respectively. When the camera focuses on the tennis ball, basketball is blurred and vice versa. In the second target light fields, the foremost person is standing on the position of floated front panel while the streetlamp is located on the rear panel. When the foremost man is focused, the streetlamp is blurred and vice versa.

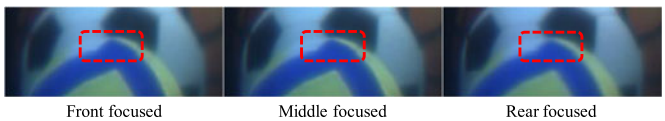


Fig. 12. Experimental results when focusing is on the front, middle, and rear positions. The upper end of the volley ball is detected clearly when the camera focuses on the middle position.

Although the system is free from the diffraction effect in pixelated structures, the system still has the resolution limits since a single projector has to focus on scattering polarizers at different depths. In the system, the gap between two scattering polarizers is 4 mm and the projection distance is 234 mm. As the gap between two scattering planes increases, the images at each plane will be out of focus significantly. To show the validity of using a single projector, it is necessary to analyze the relationship between image quality and the gap between two scattering polarizers.

There are many methods to demonstrate the image quality of display systems. We adopted slanted-edge modulation transfer function (MTF) method to measure the spatial frequency response for various projection distances [25]. Another advantage of getting MTF curve is that it also indicates the quality of scattering polarizer and projector as display devices. Slanted-edge samples are taken with varying the distance between the scattering polarizer and the projector while the focus distance is fixed to 234 mm which is same as the system condition. The distance between the projector and the scattering polarizer varies from 205 mm to 261 mm with 1 mm interval. From the experiments, 57 slanted-edge samples are achieved and slanted-edge MTF algorithm is applied to each of slanted-edge sample. Thus, 57 different MTF curves are derived. Then maximum spatial frequency values of 50%-MTF are extracted for each of MTF

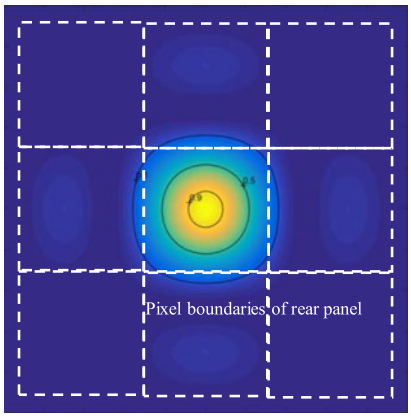


Fig. 13. Simulation for analysis of diffraction effect. As the rectangular aperture at the floated front panel is illuminated, the intensity pattern appears at the floated rear panel. Intensity pattern and pixel boundaries of the floated front panel for the experimental conditions are shown.

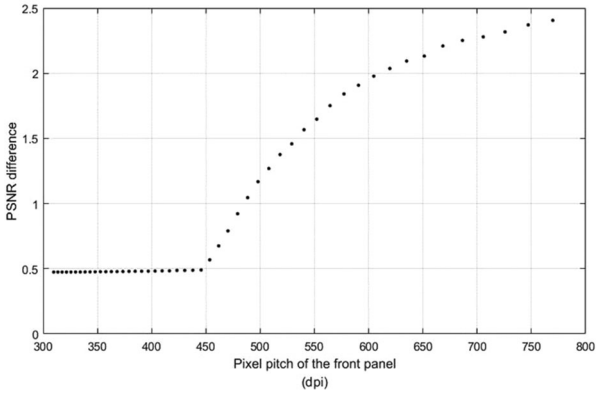


Fig. 14. PSNR difference between reconstructed images at scattering layers (diffraction-free) and LC layers (with diffraction effect). As the resolution increases, image degradation due to the diffraction effect is intensified.

curves. Fig. 15 represents the example of slanted-edge MTF algorithm for the projection distance of 234 mm. Fig. 15(a) and (b) represent the slanted image at the scattering polarizer and detected edge by the algorithm, respectively. Fig. 15(c) shows the line spread function for the given slanted edge. By applying Fourier transform, it is possible to get MTF curve as shown in Fig. 15(d).

Fig. 16 represents the maximum spatial frequency for 0.5-MTF value as varying the distance between the projector and the scattering polarizer. Since the projector is focused on 234 mm, the peak value of 17.619 appears at 234 mm. The spatial frequencies at two planes which are 2 mm away from the focus plane (232, 236 mm) are 17.466 and 17.53. Thus it is reasonable to claim that 4 mm gap is small enough to represent sharp images when a single projector is focused on somewhere between two planes. Red dashed line in Fig. 16 means 90% of peak value. If 90% of maximum spatial frequency is reasonable to present images, the gap between two scattering polarizers can increase until 15 mm. For this condition, two planes are floated at 874 mm and 152 mm, respectively.

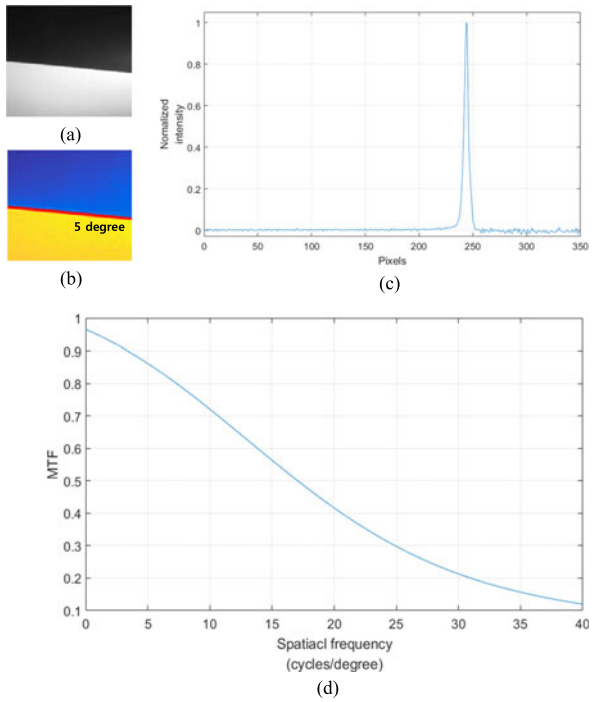


Fig. 15. Slanted-edge MTF algorithm at projection distance of 234 mm. (a) The slanted image presented at the scattering polarizer and (b) detected slanted-edge by the algorithm. (c) The line spread function for the given slanted edge and (d) MTF curve at projection distance of 234 mm.

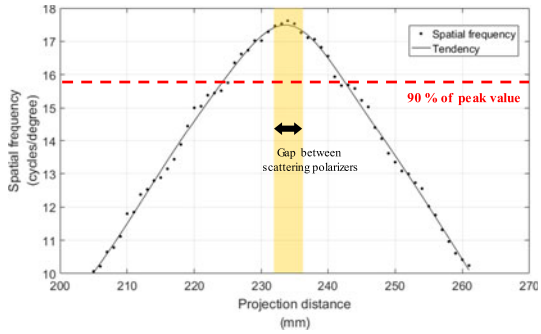


Fig. 16. The maximum spatial frequency for 0.5-MTF value for various projection distance. The peak value is 17.619 which appears at the exact focus plane. Spatial frequencies at two planes which are 2 mm away from the focus plane are 17.466 and 17.53, respectively.

IV. Conclusion

Providing visually comfortable experiences is crucial to realize the practical VR systems. One of the most important depth cues to support the visual comfort is accommodation cue. The proposed system realizes the accommodation cue by using layered scattering polarizers combined with concept of additive compressive light field display. The system generates several viewpoints to each of observer's eyes to mitigate the VAC. Since viewpoints are not fixed to a single position, the system has a potential to be adopted in VR system for people with various interpupillary gaps. Furthermore, different from stacking LCD panels, layered scattering polarizers are relatively free from diffraction effect. It presents the possibility to overcome the hindrance for enhancing spatial resolution of display devices.

At last, since the system employs only a single projector, it is easy to construct the system with compact form factor and low cost. We expect that the proposed system can overcome the vergence-accommodation conflict problem without image degradation from diffraction effect. This will make the system applied in various VR systems.

REFERENCES

[1] M. T. Schultheis, J. Himelstein, and A. A. Rizzo, "Virtual reality and neuropsychology: Upgrading the current tools," *J. Head. Trauma Rehabil.*, vol. 17, no. 5, pp. 378–394, 2002.

[2] R. Ter Haar, "Virtual reality in the military: Present and future," Presented at *3rd Twente Student Conf. IT*, 2005. [Online]. Available: http://referaat.cs.utwente.nl/conference/3/

[3] B. Lee, "Three-dimensional displays, past and present," *Phys. Today*, vol. 66, no. 4, pp. 36–41, 2013.

[4] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks, "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," *J. Vis.*, vol. 8, no. 3, 2008, Art. no. 33.

[5] J.-H. Lee, J. Park, D. Nam, S. Y. Choi, D.-S. Park, and C. Y. Kim, "Optimal projector configuration design for 300-Mpixel multi-projection 3D display," *Opt. Express*, vol. 21, no. 22, pp. 26820–26835, 2013.

[6] Y. Takaki and N. Nago, "Multi-projection of lenticular displays to construct a 256-view super multi-view display," *Opt. Express*, vol. 18, no. 9, pp. 8824–8835, 2010.

[7] C.-K. Lee, S.-G. Park, S. Moon, and B. Lee, "Viewing zone duplication of multi-projection 3D display system using uniaxial crystal," *Opt. Express*, vol. 24, no. 8, pp. 8458–8470, 2016.

[8] K. Akeley, S. J. Watt, A. R. Girshick, and M. S. Banks, "A stereo display prototype with multiple focal distances," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 804–813, 2004.

[9] S. Liu and H. Hua, "A systematic method for designing depth-fused multi-focal plane three-dimensional displays," *Opt. Express*, vol. 18, no. 11, pp. 11562–11573, 2010.

[10] S.-G. Park, J. Y. Hong, C. K. Lee, and B. Lee, "Real-mode depth-fused display with viewer tracking," *Opt. Express*, vol. 23, no. 20, pp. 26710–26722, 2015.

[11] D. Lanman, M. Hirsch, Y. Kim, and R. Raskar, "Content-adaptive parallax barriers: optimizing dual-later 3D displays using low-rank light field factorization," *ACM Trans. Graph.*, vol. 29, no. 6, 2010, Art. no. 163.

[12] G. Wetzstein, D. Lanman, W. Heidrich, and R. Raskar, "Layered 3D: Tomographic image synthesis for attenuation-based light field and high dynamic range displays," *ACM Trans. Graph.*, vol. 30, no. 4, 2011, Art. no. 95.

[13] D. Lanman, G. Wetzstein, M. Hirsch, W. Heidrich, and R. Raskar, "Polarization fields: dynamic light field display using multi-layer LCDs," *ACM Trans. Graph.*, vol. 30, no. 6, 2011, Art. no. 186.

[14] G. Wetzstein, D. Lanman, W. Heidrich, and R. Raskar, "Tensor display: Compressive light field synthesis using multilayer display with directional backlighting," *ACM Trans. Graph.*, vol. 31, pp. 1–11, 2012.

[15] M. Hirsch, G. Wetzstein, and R. Raskar, "A compressive light field projection system," *ACM Trans. Graph.*, vol. 33, no. 4, 2014, Art. no. 58.

[16] H. Fu-Chung, K. Chen, and G. Wetzstein, "The light field stereoscope: Immersive computer graphics via factored near-eye light field displays with focus cues," *ACM Trans. Graph.*, vol. 34, no. 4, 2015, Art. no. 60.

[17] S. Lee, C. Jang, S. Moon, J. Cho, and B. Lee, "Additive light field displays: Realization of augmented reality with holographic optical elements," *ACM Trans. Graph.*, vol. 35, no. 4, 2016, Art. no. 60.

[18] C.-K. Lee, S. Moon, S. Lee, D. Yoo, J.-Y. Hong, and B. Lee, "Compact three-dimensional head-mounted display system with Savart plate," *Opt. Express*, vol. 24, no. 17, pp. 19531–19544, 2016.

[19] B. E. Saleh and M. C. Teich, "Wave optics," in *Fundamentals of Photonics*, 3rd ed. Hoboken, NJ, USA: Wiley-Interscience, 1991.

[20] B. Keck, H. Hofmann, H. Scherl, M. Kowarschik, and J. Hornegger, "GPU-accelerated SART reconstruction using the CUDA programming environment," *Proc. SPIE*, vol. 7258, 2009, Art. no. 72582B.

[21] LC-Tec Displays AB, Borlänge, Sweden. [Online]. Available: http://www.lc-tec.se/

[22] K. Hong, J. Hong, J.-H. Jung, J.-H. Park, and B. Lee, "Rectification of elemental image set and extraction of lens lattice by projective image transformation in integral imaging," *Opt. Express*, vol. 18, no. 11, pp. 12002–12016, 2010.

[23] L. Ma, Y. Chen, and K. L. Moore, "Rational radial distortion models of camera lenses with analytical solution for distortion correction," *Int. J. Inform. Acquisitionc*, vol. 1, no. 2, pp. 135–147, 2004.

[24] J. W. Goodman, "Fresnel diffraction," in *Fourier Optics*, 3rd ed. Englewood, CO, USA: Roberts & Company, 2005.

[25] J. K. Roland, "A study of slanted-edge MTF stability and repeatability," *Proc. SPIE*, vol. 9396, 2015, Art. no. 93960L.



Seokil Moon received the B.S. degree in electrical engineering from Pohang University of Science and Technology, Pohang, South Korea. He is currently working toward the Ph.D. degree in the School of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea. His current research interest focuses on light field imaging technique and visualization.



Chang-Kun Lee received the B.S. and M.S. degrees in the Department of Infomation Display, Kyung Hee University, Seoul, South Korea, in 2011 and 2013, respectively. He is currently working toward the Ph.D. degree in the School of Electrical and computer Engineering, Seoul National University, Seoul, South Korea. His research interest focuses on 3-D imaging and the advanced display system.



Dukho Lee received the B.S. degree in electrical engineering, in 2016, from Seoul National University, Seoul, South Korea, where he is currently working toward the Ph.D. degree in the School of Electrical and Computer Engineering. His current research interests include the 3-D display, 3-D imaging and, frontal-projection type 3-D display.



Changwon Jang received the B.S. degree in electrical engineering, in 2013, from Seoul National University, Seoul, Korea, where he is currently working toward the Ph.D. degree in the School of Electrical and computer Engineering. His primary research interest focuses on the areas of 3-D display and digital holography.



Byoungho Lee (M'94–SM'00–F'14) received the Ph.D. degree from the Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, CA, USA, in 1993. Since September 1994, he has been in the faculty of the School of Electrical Engineering, Seoul National University, where he is currently the Head of the Department. He received several distinguished awards such as the Scientist of the Month Award of Korea (September 2009), the Academic Award of the Optical Society of Korea (2006), the Academic Award of Seoul National University (2013), and the Jinbojang Presidential Badge (2016). He was the Director-at-Large of OSA, the Chair of the Member and Education Services Council of OSA, and the Chair of the Holography and Diffractive Optics Technical Group of OSA. He has served on the editorial board of *Optics Letters*, *Applied Optics*, and *Japanese Journal of Applied Physics* as well as the Editor-in-Chief of the *Journal of the Optical Society of Korea*. He was the Vice-President of the Korean Information Display Society. He is currently the Vice-President of the Optical Society of Korea, a Fellow of the OSA and the SPIE, and a member of the Korean Academy of Science and Technology. He is on the editorial board of *Advances in Optics and Photonics, Light: Science and Applications*, and *Applied Physics B*. He is also the Editor-in-Chief of *Journal of Information Display*.

Information for Authors

(Updated/Effective October 2015)

For Transactions and Journals:

Authors are encouraged to submit manuscripts of Regular papers (papers which provide a complete disclosure of a technical premise), or Comment Correspondences (brief items that provide comment on a paper previously published in the TRANSACTIONS).

Submissions/resubmissions must be previously unpublished and may not be under consideration elsewhere.

Every manuscript must:

1. provide a clear statement of the problem and what the contribution of the work is to the relevant research community;
2. state why this contribution is significant (what impact it will have);
3. provide citation of the published literature most closely related to the manuscript; and
4. state what is distinctive and new about the current manuscript relative to these previously published works.

By submission of your manuscript to this Transaction, all listed authors have agreed to the authorship list and all the contents and confirm that the work is original and that figures, tables and other reported results accurately reflect the experimental work. In addition, the authors all acknowledge that they accept the rules established for publication of manuscripts, including agreement to pay all overlength page charges, color charges, and any other charges and fees associated with publication of the manuscript. Such charges are not negotiable and cannot be suspended. The corresponding author is responsible for obtaining consent from all co-authors and, if needed, from sponsors before submission.

In order to be considered for review, a paper must be within the scope of the journal and represent a novel contribution. A paper is a candidate for an Immediate Rejection if it is of limited novelty, e.g. a straightforward combination of theories and algorithms that are well established and are repeated on a known scenario. Experimental contributions will be rejected without review if there is insufficient experimental data. The transactions are published in English. Papers that have a large number of typographical and/or grammatical errors will also be rejected without review.

In addition to presenting a novel contribution, acceptable manuscripts must describe and cite related work in the field to put the contribution in context. Do not give theoretical derivations or algorithm descriptions that are easily found in the literature; merely cite the reference.

New and revised manuscripts should be prepared following the "Manuscript Submission" guidelines below, and submitted to the online manuscript system, ScholarOne Manuscripts. Do not send original submissions or revisions directly to the Editor-in-Chief or Associate Editors; they will access your manuscript electronically via the Manuscript Central system.

Manuscript Submission. Please follow the next steps.

1. *Account in ScholarOne Manuscripts.* If necessary, create an account in the on-line submission system ScholarOne Manuscripts. Please check first if you already have an existing account which is based on your e-mail address and may have been created for you when you reviewed or authored a previous paper.
   - All IEEE journals require an Open Researche and Contributor ID (ORCID) for all authors. ORCID is a persistent unique identifier for researchers and functions similarly to an article's Digital Object Identifier (DOI). The author will need a registered ORCID in order to submit a manuscript or review a proof in this journal.
2. *Electronic Manuscript*. Prepare a PDF file containing your manuscript in double-column, single-spaced format using a font size of 10 points or larger, having a margin of at least 1 inch on all sides. Upload this version of the manuscript as a PDF file "double.pdf" to the ScholarOne-Manuscripts site. Since many reviewers prefer a larger font, you are strongly encouraged to also submit a single-column, double-spaced version (11 point font or larger), which is easy to create with the templates provided **IEEE Author Digital Toolbox (http://www.ieee.org/publications_standards/publications/authors/authors_journals.html**). Page length restrictions will be determined by the double-column version. Proofread your submission, confirming that all figures and equations are visible in your document before you "SUBMIT" your manuscript. Proofreading is critical; once you submit your manuscript, the manuscript cannot be changed in any way. You may also submit your manuscript as a .PDF or MS Word file. The system has the capability of converting your files to PDF, however it is your responsibility to confirm that the conversion is correct and there are no font or graphics issues prior to completing the submission process.
3. *EDICS (Not applicable to Journal of Selected Topics in Signal Processing)*. All submissions must be classified by the author with an EDICS (Editors' Information Classification Scheme) selected from the list of EDICS published online at the at the publication's EDICS webpage (*please see the list below). Upon submission of a new manuscript, please choose the EDICS categories that best suit your manuscript. Failure to do so will likely result in a delay of the peer review process.
4. *Additional Documents for Review*. Please upload pdf versions of all items in the reference list that are not publicly available, such as unpublished (submitted) papers. Graphical abstracts and supplemental materials intended to appear with the final paper (see below) must also be uploaded for review at the time of the initial submission for consideration in the review process. Use short filenames without spaces or special characters. When the upload of each file is completed, you will be asked to provide a description of that file.
5. *Supplemental Materials*. IEEE Xplore can publish multimedia files (audio, images, video, pseudocode and detailed algebraic manipulations of proofs), datasets, and software (e.g. Matlab code) along with your paper. Alternatively, you can provide the links to such files in a README file that appears on Xplore along with your paper. For details, please see IEEE Author Digital Toolbox (http://www.ieee.org/publications_standards/publications/authors/authors_journals.html) under "Multimedia." To make your work reproducible by others, the TRANSACTIONS encourages you to submit all files that can recreate the figures in your paper.
6. *Submission*. After uploading all files and proofreading them, submit your manuscript by clicking "Submit." A confirmation of the successful submission will open on screen containing the manuscript tracking number and will be followed with an e-mail confirmation to the corresponding and all contributing authors. Once you click "Submit," your manuscript cannot be changed in any way.
7. *Copyright Form and Consent Form*. By policy, IEEE owns the copyright to the technical contributions it publishes on behalf of the interests of the IEEE, its authors, and their employers; and to facilitate the appropriate reuse of this material by others. To comply with the IEEE copyright policies, authors are required to sign and submit a completed "IEEE Copyright and Consent Form" prior to publication by the IEEE. The IEEE recommends authors to use an effective electronic copyright form (eCF) tool within the ScholarOne Manuscripts system. You will be redirected to the "IEEE Electronic Copyright Form" wizard at the end of your original submission; please simply sign the eCF by typing your name at the proper location and click on the "Submit" button.

Comment Correspondence. Comment Correspondences provide brief comments on material previously published in these TRANSACTIONS. These items may not exceed 2 pages in double-column, single spaced format, using 9 point type, with margins of 1 inch minimum on all sides, and including: title, names and contact information for authors, abstract, text, references, and an appropriate number of illustrations and/or tables. Correspondence items are submitted in the same way as regular manuscripts (see "Manuscript Submission" above for instructions).

Authors may also submit manuscripts of overview articles, but note that these include an additional white paper approval process http://www.signalprocessingsociety.org/publications/overview-articles/. *[This does not apply to the Journal of Selected Topics in Signal Processing. Please contact the Editor-in-Chief.]*

Digital Object Identifier 10.1109/JSTSP.2017.2761224

Manuscript Length. For the initial submission of a regular paper, the manuscript may not exceed 13 double-column pages (10 point font), including title; names of authors and their complete contact information; abstract; text; all images, figures and tables, appendices and proofs; and all references. Supplemental materials and graphical abstracts are not included in the page count. For regular papers, the revised manuscript may not exceed 16 double-column pages (10 point font), including title; names of authors and their complete contact information; abstract; text; all images, figures and tables, appendices and proofs; and all references. For Overview Papers, the maximum length is double that for regular submissions at each stage (please reference http://www.signalprocessingsociety.org/publications/overview-articles/ for more information).

Note that any paper in excess of 10 pages will be subject to mandatory overlength page charges. Since changes recommended as a result of peer review may require additions to the manuscript, it is strongly recommended that you practice economy in preparing original submissions. Note: Papers submitted to the TRANSACTIONS ON MULTIMEDIA in excess of 8 pages will be subject to mandatory overlength page charges.

Exceptions to manuscript length requirements may, under extraordinary circumstances, be granted by the Editor-in-Chief. However, such exception does not obviate your requirement to pay any and all overlength or additional charges that attach to the manuscript.

Resubmission of Previously Rejected Manuscripts. Authors of manuscripts rejected from any journal are allowed to resubmit their manuscripts only once. At the time of submission, you will be asked whether your manuscript is a new submission or a resubmission of an earlier rejected manuscript. If it is a resubmission of a manuscript previously rejected by any journal, you are expected to submit supporting documents identifying the previous submission and detailing how your new version addresses all of the reviewers' comments. Papers that do not disclose connection to a previously rejected paper or that do not provide documentation as to changes made may be immediately rejected.

Author Misconduct. Author misconduct includes plagiarism, self-plagiarism, and research misconduct, including falsification or misrepresentation of results. All forms of misconduct are unacceptable and may result in sanctions and/or other corrective actions. Plagiarism includes copying someone else's work without appropriate credit, using someone else's work without clear delineation of citation, and the uncited reuse of an author's previously published work that also involves other authors. Self-plagiarism involves the verbatim copying or reuse of an author's own prior work without appropriate citation, including duplicate submission of a manuscript elsewhere (journal or conference), and submission of two different manuscripts which overlap substantially in language or technical contribution. The aforementioned forms of author misconduct also cover materials from third party servers (such as Arxiv). For more information on the definitions, investigation process, and corrective actions related to author misconduct, see the Signal Processing Society Policies and Procedures Manual, Section 6.1. Author misconduct may also be actionable by the IEEE under the rules of Member Conduct.

Extensions of the Author's Prior Work. It is acceptable for conference papers to be used as the basis for a more fully developed journal publication. However, authors are required to cite their related prior work, either in the introduction or in a footnote. The papers cannot be identical, and the journal paper should be justified by a clearly identifiable benefit that its publication offers to the research community beyond the already published conference paper. For example, the journal paper may include additional analysis, novel algorithmic enhancements, added theoretical work, completeness of exposition, extensive experimental validation, etc. The added benefits of the journal paper must either be apparent from a reading of the introduction or abstract, or should be clearly and concisely explained in a separate document that accompanies the submission.

Submission Format. Authors are required to prepare manuscripts employing the on-line style files developed by IEEE, which include guidelines for abbreviations, mathematics, and graphics. All manuscripts accepted for publication will require the authors to make final submission employing these style files. The style files are available on the web at the **IEEE Author Digital Toolbox** under "Template for all TRANSACTIONS." (LaTeX and MS Word). Please note the following requirements about the abstract:

- The abstract must be a concise yet comprehensive reflection of what is in your article.
- The abstract must be self-contained, without abbreviations, footnotes, displayed equations, or references.
- The abstract must be between 150-250 words.
- The abstract should include a few keywords or phrases, as this will help readers to find it. Avoid over-repetition of such phrases as this can result in a page being rejected by search engines.

In addition to written abstracts, papers may include a graphical abstract; see http://www.ieee.org/publications_standards/publications/authors/authors_journals.html. for options and format requirements.

IEEE supports the publication of author names in the native language alongside the English versions of the names in the author list of an article. For more information, see "Author names in native languages" (http://www.ieee.org/publications_standards/publications/authors/auth_names_native_lang.pdf) on the IEEE Author Digital Toolbox page.

Refining the Use of English Language in Your Manuscript. A professional editing service, American Journal Experts (AJE) [http://www.aje.com/en] is available to authors looking to refine and polish the content of their papers for a fee. You are also free to select another professional editing service, or to ask a colleague who is fluent in English to assist with editing. However, if the revised manuscript does not meet the English usage criteria, then you must use the designated editing service (AJE [http://www.aje.com/en]) for a fee, or you must withdraw the manuscript.

Open Access. The publication is a hybrid journal, allowing either Traditional manuscript submission or Open Access (author-pays OA) manuscript submission. Upon submission, if you choose to have your manuscript be an Open Access article, you commit to pay the discounted $1,950 OA fee if your manuscript is accepted for publication in order to enable unrestricted public access. Any other application charges (such as overlength page charge and/or charge for the use of color in the print format) will be billed separately once the manuscript formatting is complete but prior to the publication. If you would like your manuscript to be a Traditional submission, your article will be available to qualified subscribers and purchasers via IEEE Xplore. No OA payment is required for Traditional submission.

Page Charges.

Voluntary Page Charges. Upon acceptance of a manuscript for publication, the author(s) or his/her/their company or institution will be asked to pay a charge of $110 per page to cover part of the cost of publication of the first ten pages that comprise the standard length (two pages, in the case of Correspondences).

Mandatory Page Charges. The author(s) or his/her/their company or institution will be billed $220 per each page in excess of the first ten published pages for regular papers. (**NOTE: Regular papers accepted to IEEE TRANSACTIONS ON MULTIMEDIA in excess of 8 pages will be subject to mandatory overlength page charges, and correspondence papers accepted to T-MM in excess of 6 published pages will also be subject to overlength page charges.) These are mandatory page charges and the author(s) will be held responsible for them. They are not negotiable or voluntary. The author(s) signifies his willingness to pay these charges simply by submitting his/her/their manuscript to the TRANSACTIONS. The Publisher holds the right to withhold publication under any circumstance, as well as publication of the current or future submissions of authors who have outstanding mandatory page charge debt. No mandatory overlength page charges will be applied to overview articles in the Society's journals.

Color Charges. Figures may appear in color, in the electronic (Xplore) version only, free of charge. The figure will be printed in the hardcopy version in grayscale and the author is responsible for ensuring that the corresponding grayscale figure is understandable. Color reproduction charges for print are the responsibility of the author. Details of the associated charges can be found on the IEEE Publications page.

Payment of fees on color reproduction is not negotiable or voluntary, and the author's agreement to publish the manuscript in these TRANSACTIONS is considered acceptance of this requirement.

*EDICS Webpages:

IEEE TRANSACTIONS ON SIGNAL PROCESSING:
http://www.signalprocessingsociety.org/publications/periodicals/tsp/TSP-EDICS/
IEEE TRANSACTIONS ON IMAGE PROCESSING:
http://www.signalprocessingsociety.org/publications/periodicals/image-processing/tip-edics/
IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE / ACM:
http://www.signalprocessingsociety.org/publications/periodicals/taslp/taslp-edics/
IEEE TRANSACTIONS ON INFORMATION, FORENSICS AND SECURITY:
http://www.signalprocessingsociety.org/publications/periodicals/forensics/forensics-edics/
IEEE TRANSACTIONS ON MULTIMEDIA:
http://www.signalprocessingsociety.org/tmm/tmm-edics/
IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING:
http://www.signalprocessingsociety.org/publications/periodicals/tci/tci-edics/
IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS:
http://www.signalprocessingsociety.org/publications/periodicals/tsipn/tsipn-edics/

IEEE Collabratec™
Bright Minds. Bright Ideas.



Introducing IEEE Collabratec™

The premier networking and collaboration site for technology professionals around the world.

IEEE Collabratec is a new, integrated online community where IEEE members, researchers, authors, and technology professionals with similar fields of interest can **network** and **collaborate**, as well as **create** and manage content.

Featuring a suite of powerful online networking and collaboration tools, IEEE Collabratec allows you to connect according to geographic location, technical interests, or career pursuits.

You can also create and share a professional identity that showcases key accomplishments and participate in groups focused around mutual interests, actively learning from and contributing to knowledgeable communities. All in one place!

Network.
Collaborate.
Create.

Learn about IEEE Collabratec at
ieeecollabratec.org

IEEE

Digital Object Identifier 10.1109/JSTSP.2017.2773840

IEEE Access®

The **journal** for rapid **open access** publishing

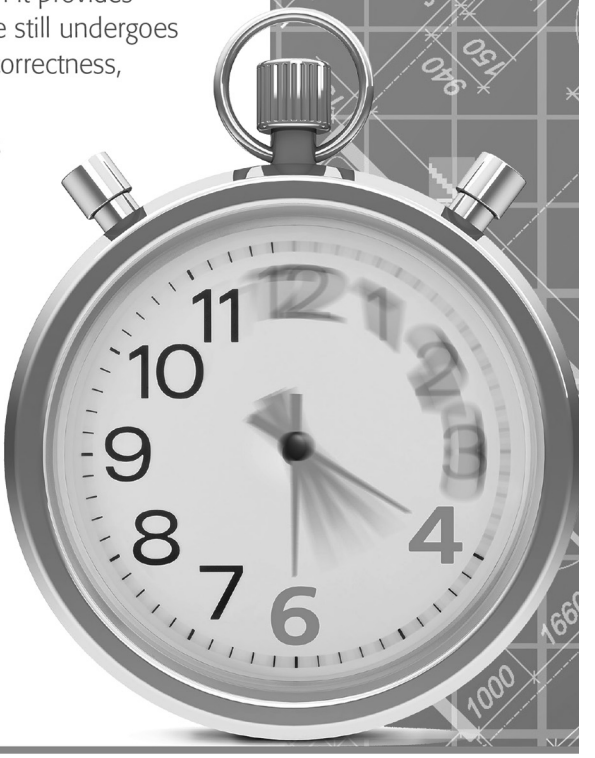
Best New Journal in STM 2015

Become a published author in 4 to 6 weeks.

Get on the fast track to publication with the multidisciplinary open access journal worthy of the IEEE.

IEEE journals are trusted, respected, and rank among the most highly cited publications in the industry. IEEE *Access* is no exception with a typical **one-third** acceptance rate. Even though it provides authors faster publication time, every submitted article still undergoes extensive peer review to ensure originality, technical correctness, and interest among readers.

Published only online, IEEE *Access* is ideal for authors who want to quickly announce recent developments, methods, or new products to a global audience.



Publishing in IEEE *Access* allows you to:

- Submit multidisciplinary articles that do not fit neatly in traditional journals
- Reach millions of global users through the IEEE *Xplore*® digital library with free access to all
- Integrate multimedia with articles
- Connect with your readers through commenting
- Track usage and citation data for each published article
- Publish without a page limit for **only $1,750** per article

Learn more about this award-winning journal at:
www.ieee.org/ieee-access

IEEE
Advancing Technology for Humanity

14-PUB-196 11/15

Digital Object Identifier 10.1109/JSTSP.2017.2773842

IEEE Membership Can Help You **Reach Your Personal and Professional Goals**



Gain access to the latest IEEE news, publications and digital library. Give and receive personal mentoring. Network locally and globally with IEEE members. And that's only the beginning. Discover how IEEE can help jumpstart your career.

"Participating in IEEE has developed me as a well-rounded engineer and helped me shine during networking events."

-Likhitha Patha
Electrical Engineering Student,
IEEE Brand President,
Virginia Polytechnic Institute
and State University

Visit **www.ieee.org/join** today.

IEEE

Digital Object Identifier 10.1109/JSTSP.2017.2773844

IEEE SIGNAL PROCESSING SOCIETY

Board of Governors

R. WARD, *President*
University of British Columbia
Vancouver BC V6T 1Z4 Canada

A. SAYED, *President-Elect*
University of California
Los Angeles, CA

N. SIDIROLOULOS, *Vice President–Membership*
University of Minnesota
Minneapolis, MN 55455

C. REGAZZONI, *Vice President-Conferences*
University of Genova
Genova, Italy

T. PAPPAS, *Vice President Publications*
Northwestern University
Evanston, IL 60208-3118

W. KELLERMANN, *Vice President-Technical Directions*
Lehrstuhl für Multimediakommunikation und Signalverarbeitung
Cauerstraße 7
91058 Erlangen

H. VINCENT POOR, *Awards Board Chair*
Princeton University
Princeton, NJ 08544

R. BASEIL, *Executive Director*
Piscataway, NJ 08854

Members-at-Large
Term Ending 31 December 2017
M. GRECO
S. YOUNG
A. ZOUBIR

Members-at-Large
Term Ending 31 December 2018
R. W. HEATH Jr.
L. J. KARAM
M. WU

Members-at-Large
Term Ending 31 December 2019
A. A. H. ALWAN
A. ORTEGA
B. PESQUET-POPESCU

Regional Director-at-Large
Term Ending 31 December 2017
(Regions 7 & 9) P. S. R. DINIZ
(Region 10) H. KIYA

Regional Director-at-Large
Term Ending 31 December 2018
(Regions 1 & 6) Z. WANG
(Region 8) A. I. PÉREZ-NEIRA

Standing Committee Chairs
Chapters, A. MAJUMDAR
Fellow Reference, I. TRANCOSO
Industrial Relations, P. NASIOPOULOS
Long-Range Planning and Implementation, A. SAYED
Membership Services, K. LAM
Student Services, P. CAMPISI
Nominations and Appointments, A. ACERO
TC Review Committee, A. SAYED
Chapters Review, SIDIROPOULOS
Education, M. BUGALLO
Regional 1–6, Z. WANG
Regional 7 & 9, P. S. R. DINIZ
Regional 8, A. I. PÉREZ-NEIRA
Regional 10, H. KIYA
Resource Center, J. MCALLISTER
Sig Port, Y. SUN
Seasonal Schools SubCommittee & Chair, C.-T. CHIU
Women in Signal Processing Subcommittee, A. PAPANDREOU-SUPPAPPOLA
Young Professionals Subcommittee, M. POURAZAD

Technical Committee Chairs
Audio and Acoustics Signal Processing, S. GANNOT
Bio Imaging and Signal Processing, A. MUÑOZ-BARRUTIA
Design and Implementation of Signal Processing Systems, Z. YAN
Image, Video, and Multidimensional Signal Processing, B. PESQUET-POPESCU
Industry DSP Technology Standing Committee, F.-L. LUO
Information Forensics and Security, P. CAMPISI
Machine Learning for Signal Processing, R. RAICH
Multimedia Signal Processing, E. MAGLI
Sensor Array and Multichannel, M. HAARDT
Signal Processing for Communications and Networking, W. YU
Signal Processing Theory and Methods, A.-J. VAN DER VEEN
Speech and Language Processing, M. BACCHIANI

Awards Board
Chair, H. VINCENT POOR

Conference Board
Chair, C. REGAZZONI

Membership Board
Chair, SIDIROPOULOS

Publications Board
Chair, T. PAPPAS

Technical Directions Board
Chair, W. KELLERMANN

Publications Editors-in-Chief

Signal Processing Letters
J. FOWLER
Mississippi State University
Mississippi State, MS 39762

Signal Processing Magazine
M. WU
University of Maryland
College Park, MD 20742

Trans. Signal Processing
S. THEODORIDIS
University of Athens
Greece

Trans. Information Forensics and Security
M. BARNI
Universita di Siena
Italy

Trans. Image Processing
S. ACTON
University of Virginia
Charlottesville, VA 22902

IEEE/ACM Trans. Audio, Speech, and Language Processing
H. LI
Institute for Infocomm Research
Singapore

Journal Selected Topics in Signal Processing
S. NARAYANAN
University of Southern California
Los Angeles, CA

SPS Publications Office
IEEE Signal Processing Society
Piscataway, NJ 08854
sp.pub.info@ieee.org

K. JACKSON, *Coordinator Society Publications*
IEEE Transactions on Image Processing, IEEE/ACM Transactions on Audio, Speech and Language Processing

R. WOLLMAN, *Publications Administrator*
IEEE Signal Processing Magazine, IEEE Signal Processing Letters

A. FISHER, *Coordinator Society Publications*
IEEE Transactions on Signal Processing, IEEE Transactions on Computational Imaging, Transactions on Multimedia

A. Fleisher, *Coordinator Society Publications*
IEEE Transactions on Information Forensics and Security, IEEE Journal of Selected Topics in Signal Processing, IEEE Transactions on Signal and Information Processing Over Networks

Liaison Representatives

AVSS Steering Committee, P. FROSSARD, H. KO, J. KONRAD, E. MONARI, S. VELIPASALAR
IEEE Biometrics Council, P. FLYNN
Brain Initiative, J. PRINCIPE
IEEE Cloud Computing, N. SIDIROPOULOS, J. DROPPO – *Conferences*
IEEE Computing in Science and Engineering Magazine, M. CHAKRABORTY
Council on Radio Frequency Identification, S. PATIL, R. ZHANG
IEEE ICME Steering Committee, D. XU, X.-P. ZHANG
IEEE Control Systems Letters, A. K. ROY CHOWDHURY
IEEE Conferences Committee, W.-C. SIU
IEEE Life Sciences Letters / Life Sciences Technical Community Initiative, A. M. BARRUTIA, M. LIEBLING
IEEE Women in Engineering Committee, S. WOOD
IEEE Internet of Things Journal Steering Committee, R. YU
IEEE Internet of Things World Forum, R. YU
IEEE ISBI Steering Committee, J. FESSLER, A. M. BARRUTIA, M. UNSER
IEEE Journal on Biomedical Health Informatics, E. FOSLER-LUSSIER, Y. J. WANG
IEEE MultiMedia Magazine, A. CAVALLARO
PSPB Author Education, T. N. PAPPAS
IEEE Security and Privacy Magazine, A. ROCHA
IEEE Sensors Council, A.-H. SAYED
Society on Social Implications of Technology, A. PIVA
IEEE Software Defined Networks (SDN) Initiative, B. EVANS
IEEE TAB Magazine Committee, M. WU
IEEE Management Committee, K. J. R. LIU
TAB Periodicals Committee, P. WILLETT
TAB Strategic Planning Committee, R. WARD
Transactions on Affective Computing, A. POTAMIANOS
Transactions on Big Data M. SLANEY
IEEE Transactions on Cognitive Communications and Networks, G. LI, S. (R.) CUI
Transactions on Computational Imaging Technical Liaison Committee, C. BOUMAN
IEEE Transactions on Medical Imaging, D. BROOKS, E. MEIJERING
IEEE Transactions on Mobile Computing Steering Committee, W. LIAO, SIDIROPOULOS
IEEE Transactions on Multimedia Steering Committee, T. PAPPAS, C. GUILLEMOT
IEEE Transactions on Signal & Information Processing Over Networks Technical Liaison Committee, A. ORTEGA
IEEE Transactions on Smart Grid / IEEE Smart Grid Initiative Steering Committee, H. ZHU
IEEE Transactions on Wireless Communications Steering Committee, U. MITRA, G. MATZ
IEEE Wireless Communications Letters Steering Committee, Y. HUA, D. (O.) WU

Digital Object Identifier 10.1109/JSTSP.2017.2761226