Marine Object Detection in UAV Full-Motion Video

Shibin Parameswaran, Corey Lane, Bryan Bagnall, Heidi Buck
Space and Naval Warfare Systems Center Pacific
53560 Hull Street, San Diego, CA 92152-5001
{shibin.parameswaran, corey.lane, bryan.bagnall, heidi.buck}@navy.mil

***Abstract*—Recent years have seen an increased use of Unmanned Aerial Vehicles (UAV) with video-recording capability for Maritime Domain Awareness (MDA) and other surveillance operations. In order for these efforts to be effective, there is a need to develop automated algorithms to process the full-motion videos (FMV) captured by UAVs in an efficient and timely manner to extract meaningful information that can assist human analysts and decision makers. This paper presents a generalizeable marine object detection system that is specifically designed to process raw video footage streaming from UAVs in real-time. Our approach does not make any assumptions about the object and/or background characteristics because, in the MDA domain, we encounter varying background and foreground characteristics such as boats, bouys and ships of varying sizes and shapes, wakes, white caps on water, glint from the sun, to name but a few. Our efforts rely on basic signal processing and machine learning approaches to develop a generic object detection system that maintains a high level of performance without making prior assumptions about foreground-background characteristics and does not experience abrupt performance degradation when subjected to variations in lighting, background characteristics, video quality, abrupt changes in video perspective, size, appearance and number of the targets. In the following report, in addition to our marine object detection system, we present representative object detection results on some real-world UAV full-motion video data.**

I. Introduction

The advent of Unmanned Aerial Vehicles (UAV) that can record and transmit live videos to base stations has presented new opportunities for improved Maritime Domain Awareness (MDA). UAVs complement traditional imagery sources such as satellite imagery or videos recorded by cameras attached to buoys, ships or harbors by leveraging their mobility to perform surveillance operations on larger areas. However, the volume and the real-time nature of the data collected by UAVs demand novel algorithms that can process and extract intelligence from full-motion videos (FMV). Unlike other applications in the video domain, the videos collected by UAVs for MDA purposes have unique characteristics making object detection a very challenging problem, namely:

- Independent motion of system components - the sensor (camera), the platform (UAV), the foreground and background objects (ship, boat, water)
- Continuous alterations of background due to waves and glint from the sun
- The presence or absence of wakes and white caps in the water
- Lack of continuity of UAV recordings resulting in abrupt scene and perspective changes
- Illumination variability
- Absence of any landmarks like building edges, trees, horizon etc.
- Medium to low quality video, with varying degrees of shakiness.

Fig. 1 illustrates some examples of frames from UAV video data.

The majority of existing object detection algorithms make assumptions about data, reducing their applicability in our problem domain which involves unconstrained and uncontrolled real-world settings. This makes our application unique and challenging, which warrants a return to the basics of detection techniques and an investigation to make them robust and generalizeable. To this end, we have developed a generic object detection system that is robust to illumination variations, ever-changing background characteristics, video quality, abrupt changes in perspective, size, appearance and number of the targets, and makes no prior assumptions about foreground-background characteristics.

The system discussed in this paper, the RAPid Image Exploitation Resource (RAPIER®) Full-Motion Video (FMV) System was developed as a follow-on to the RAPIER® Ship Detection System for satellite and still-frame imagery [1], [2], [3]. The FMV capability focuses on automatically detecting and tracking maritime targets (against a maritime background) in full-motion video; it is intended to be a video analysis tool for extracting meaningful data from FMV with minimal human supervision.

The detection and tracking of targets in FMV is currently a manual task that requires trained analysts to spend considerable amounts of time with each video. Each video must be independently watched and carefully monitored for targets. When an analyst recognizes a target, information about the target must be manually recorded. This method is time consuming, expensive, requires training, and can be error prone. Due to the immense amount of military video data produced, it is currently impractical to sufficiently analyze all of it.

There are two main objectives for this tool. The first objective is to automatically generate the first phase analysis of a video that an analyst would create. This tool can be used to quickly and automatically summarize large quantities of video, drastically reducing the considerable manpower required to complete first phase analysis of FMV data. With the aid of this tool, an analyst can spend time concentrating on important moments in the analyzed videos instead of watching each video segment all the way through. The second objective is to alert a user that an object (a ship) is present in a live maritime video

(a) Video-1, Frame-A (b) Video-1, Frame-B (c) Video 2

Fig. 1: Frames showing inter-video and intra-video variations, differences in object sizes and other factors that influence detection performance like wakes and lighting changes

feed. The system will notify the user that a new target is present and annotate the video stream by highlighting the target and displaying geographical coordinate information. Automatically identifying and alerting the observer to the presences of key targets can help relieve observation teams from having to double or triple check the feed and reduce the human error factor.

This paper provides a description of the workflow of the current implementation of RAPIER® FMV marine object detection system with an emphasis on its anomaly detection module.

II. Related Work

Much of the work in the area of object detection in video data has been carried out with specific objects in mind or with prior knowledge of background or foreground characteristics. Here, we describe a few recent research efforts in the area of object detection from aerial video imagery. Although, we share a similar goal, these investigations differ from our efforts due to the challenges of our domain and our sensor characteristics.

In [4] and [5], the authors report methodologies for the detection of people and static or moving vehicles collected from UAV imagery. Their work uses a combination of thermal imagery and optical imagery to detect people and vehicles relying on shape matching and vehicle size constraints. The work presented in [6] addresses moving vehicle detection for road traffic surveillance, by identifying the consistent movement of subsets (groups) of edges and verifying them as a moving vehicle. This depends on the assumption that it is possible to isolate moving components in a scene and that the vehicle is the only moving element of the scene. There has also been research where background stabilization is performed before attempting the detection and tracking [7]. The research reported in [8] investigates the plausibility of detection and identification of ground objects during flight. The objects detected in this work are signs/symbols made of plywood of a given size, shape or color placed on the ground. The research conducted in [14] describes a system that detects objects moving relative to their background. This system works with both the sensor and targets in motion. It differs from our system by the fact that the surface that our maritime targets move on is not static and therefore the utility of optical flow techniques is diminished.

In terms of applications in MDA domain, recent efforts have been made towards marine vehicle detection and tracking from surveillance videos. However, unlike our system, these efforts assume the existence of constant reference artifacts such as the horizon, or have constant background scenery due to stationary port-based cameras. The work in [9] addresses detection and tracking of marine vehicles in videos recorded by a camera attached to a buoy. Although the systems are in motion, this sensor location gives a view of the horizon which is used to stabilize the image prior to object detection. In the case of port-based cameras, the constancy of the background scenery (and the view of the horizon) can be leveraged by reliably modeling the background variations encountered from that point-of-view which can then be leveraged to detect foreground anomalies [10].

III. System Description

Our goal is to develop a generalizeable marine object detection framework capable of detecting and tracking objects irrespective of size, background or illumination changes. The following is a brief description of our object detection system workflow, which is also illustrated in Fig. 2.

A. Metadata Supplementation and Correction

In this stage, the metadata information available from the UAV video data is analyzed and processed to derive supplemental information that can be used to correct or substitute for any missing information in later stages for visualization or geo-location. For example, if a function for projecting image coordinates to geographic coordinates (transform matrix) is not provided with the pre-processed data, then we can deduce that from other parameters such as the sensor telemetry. This transform allows us to estimate the ground sample distance (GSD) which in turn allows us to estimate the physical dimensions of the targets. Additionally, this stage also corrects known issues with certain sensors metadata deviating from the Motion Imagery Standards Board (MISB) standards.

B. Image Pre-processing

The raw video frames are run through several preprocessing steps. Depending on the sensor, the frames might have unusable boundary pixels or burned-in metadata that will hinder the performance in the detection stage. For this

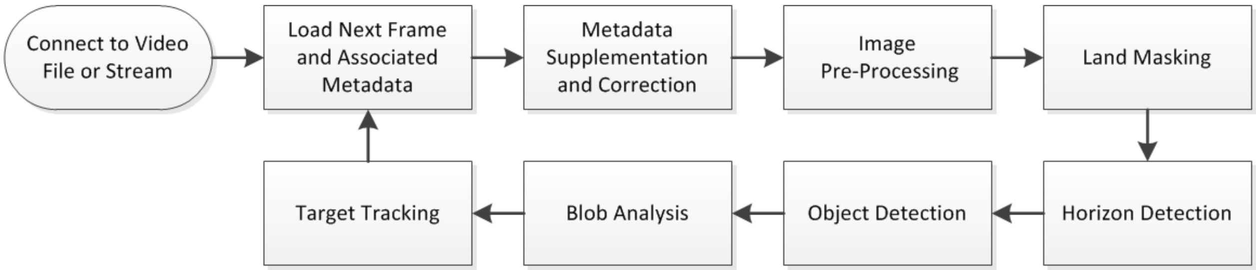

Fig. 2: Block diagram depicting the workflow of RAPIER® FMV marine object detection system

reason, each frame is cropped to select a region of interest (by discarding boundary pixels - set to 10 pixels in our studies), remove video artifacts and burned-in analog metadata from the image. Second, if the image (frame) needs enhancement due to environmental conditions (e.g., fog, haze, mist), contrast enhancement algorithms are applied. Finally, an image pyramid of the frame is generated for algorithms that may run on reduced resolution versions of the frame.

C. Land Masking

Since our system is built for marine object detection, it is important for its performance and efficiency to disable or ignore an entire frame or regions of a frame that contains land. This is called land masking. We use the World Vector Shoreline (WVS)[1] to aid our decision to mask our frame areas recorded over land. The WVS is a shapefile containing all of the known shorelines in the world which is produced by the National Geospatial-Intelligence Agency (NGA). Using this in conjunction with the telemetry from the sensor, we determine whether an area of an image is recorded over water or land. This method is very effective for disabling processing when the sensor is flying over large areas of land, but the low resolution of WVS coordinates and the ever-changing nature of shorelines over time could make this land masking solution less robust in coastal regions.

D. Horizon Detection

The majority of the captured data that we process does not contain the horizon. However, it is possible that the camera can pan upwards and capture the horizon. In scenarios where the metadata implies there may be a horizon in the frame, an algorithm will be triggered to separate the sky from the water. This, just like land masking, is used to ignore any anomalies that are detected above the horizon. Due to the drastically varying conditions in which horizons are present, unsupervised classification methods are used for statistically separating the image. Currently, a k-means clustering algorithm is used to obtain initial separation between regions, and then prior knowledge that the horizon is a straight line is used to determine a best fit line for splitting the two regions.

E. Object Detection

In the object detection stage, we use separate supervised and unsupervised methods to isolate man-made objects from the background. Currently, the algorithms that have been included in the system based on their more general applicability are an adaptive constant false alarm rate (CFAR) anomaly detection methodology and a quadratic discriminant analysis (QDA) classifier using discrete cosine transform (DCT) features.

1) Adaptive CFAR anomaly detection: Constant false alarm rate (CFAR) anomaly detection algorithm is based on the adaptive hypothesis test algorithm proposed by Yu et al. [11] which is used to detect small targets amid locally Gaussian background noise or clutter. Specifically, the algorithm works by estimating the power of the cell in question and comparing it to the local power estimates from the adjacent cells. Since the background clutter present in real-world data changes in a spatio-temporal manner, in order to achieve a constant false alarm rate, this algorithm uses a changing threshold above which a cell is flagged as anomalous. For instance, if a subset of N pixels from a local neighborhood of a J-band image is denoted by matrix $\mathbf{X} \in \mathcal{R}^{J \times N}$, and $\mathbf{b} \in \mathcal{R}^{J}$ and $\mathbf{s} \in \mathcal{R}^{N}$ denote a normalized target vector in spectral and spatial domain respectively, then the decision metric, $\lambda(\mathbf{X})$, and the threshold which is dependent on the probability of false alarm (P_{FA}), is given by

$$\lambda(\mathbf{X}) = \frac{\left(\mathbf{b^T \Sigma^{-1} X s}\right)^{\mathbf{2}}}{\left(\mathbf{b^T \Sigma^{-1} X s}\right)\left(\mathbf{1} - \frac{\mathbf{b^T X^T \Sigma^{-1} s}}{\mathbf{N}}\right)} \geq \frac{N}{\frac{1}{r_0} - 1} \tag{1}$$

where $\mathbf{\Sigma} = \frac{1}{N}\left(\mathbf{X}\mathbf{X}^T\right)$ and r_0 is estimated from

$$P_{FA} = \int_{r_0}^{1} \left[\frac{\Gamma\left(\frac{N-J+1}{2}\right)}{\Gamma\left(\frac{N-J}{2}\right)\Gamma\left(\frac{1}{2}\right)}\right](1-r)^{\frac{N-J-2}{2}}\, r^{\frac{-1}{2}} \tag{2}$$

for $0 < r < 1$

We use a fully adaptive version of CFAR, commonly known as an RX detector, which automatically estimates the spectral signature and clutter statistics with only the target's spatial information (a rough estimate of target size) [11]. The decision function under this method is given by

$$r(\mathbf{X}) = (\mathbf{X}\mathbf{s})^T \mathbf{\Sigma}^{-1} (\mathbf{X}\mathbf{s}) \geq N r_0 \tag{3}$$

2) DCT-QDA anomaly detection: A discrete cosine transform (DCT) is used to express a finite sequence of data as a sum of cosine functions with different frequencies. It is a frequency domain representation that is used extensively in signal compression due to its superior energy compaction

[1] http://shoreline.noaa.gov/data/datasheets/wvs.html

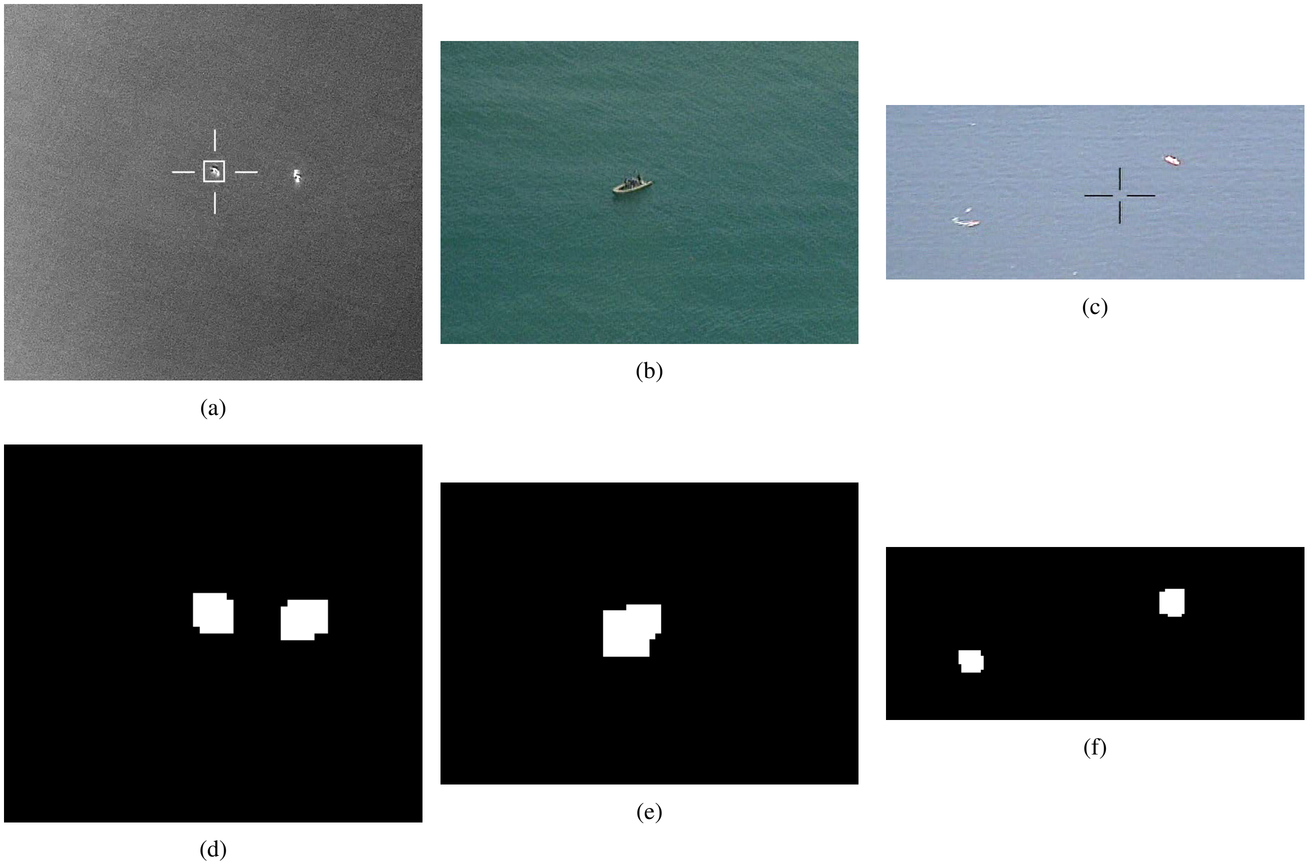

Fig. 3: Results obtained using CFAR detection algorithm. Top row (a-c) shows sample input images of CFAR detection module, and masks obtained from this anomaly detection methodology is shown in the bottom row (d-f)

property and computational efficiency [12]. For images, a two-dimensional variant of DCT (2D DCT) is used which entails running the one-dimensional DCT first along the rows and then along the columns. The 2D DCT of an $M \times N$ block of image (or simply, a matrix) is given by the following formula (ignoring scaling factors):

$$C_{uv} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x_{mn} \cos\left[(2m+1)\nu_u\right] \cos\left[(2n+1)\nu_v\right] \quad (4)$$

where, $\nu_u = \dfrac{\pi u}{2M}$ and $\nu_v = \dfrac{\pi v}{2M}$

Here, C_{uv} is the coefficient in u-th row and v-th column of 2D DCT matrix, and x_{mn} is the image pixel at spatial coordinates (m, n). To extract features for anomaly detection, the image is divided into blocks of size 8×8 pixels and 2D DCT coefficients are calculated for each block. As is the convention, the 2D DCT coefficient matrix representing the 8×8 image block is then vectorized using zigzag ordering, as shown in Fig. 5. This keeps the horizontal and vertical frequencies in increasing order and the coefficient variances in decreasing order. These feature vectors are then classified as part of background or foreground (anomaly) using a quadratic discriminant analysis classifier.

Quadratic discriminant analysis (QDA) is a variant of linear discriminant analysis (LDA) [13]. Unlike LDA, QDA does not

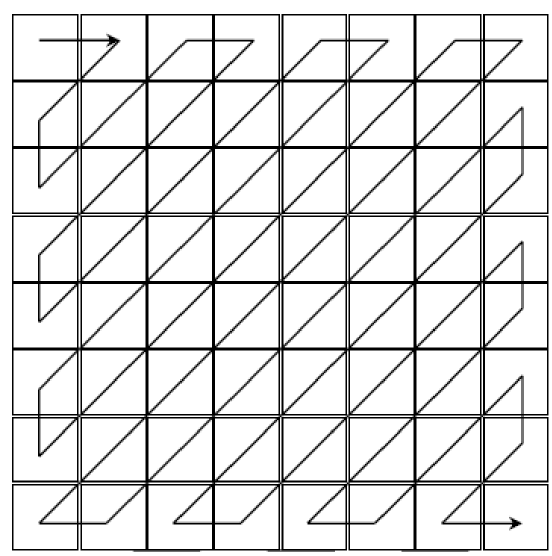

Fig. 5: Zigzag ordering used to convert 2D matrix of DCT coefficients into a 1D vector

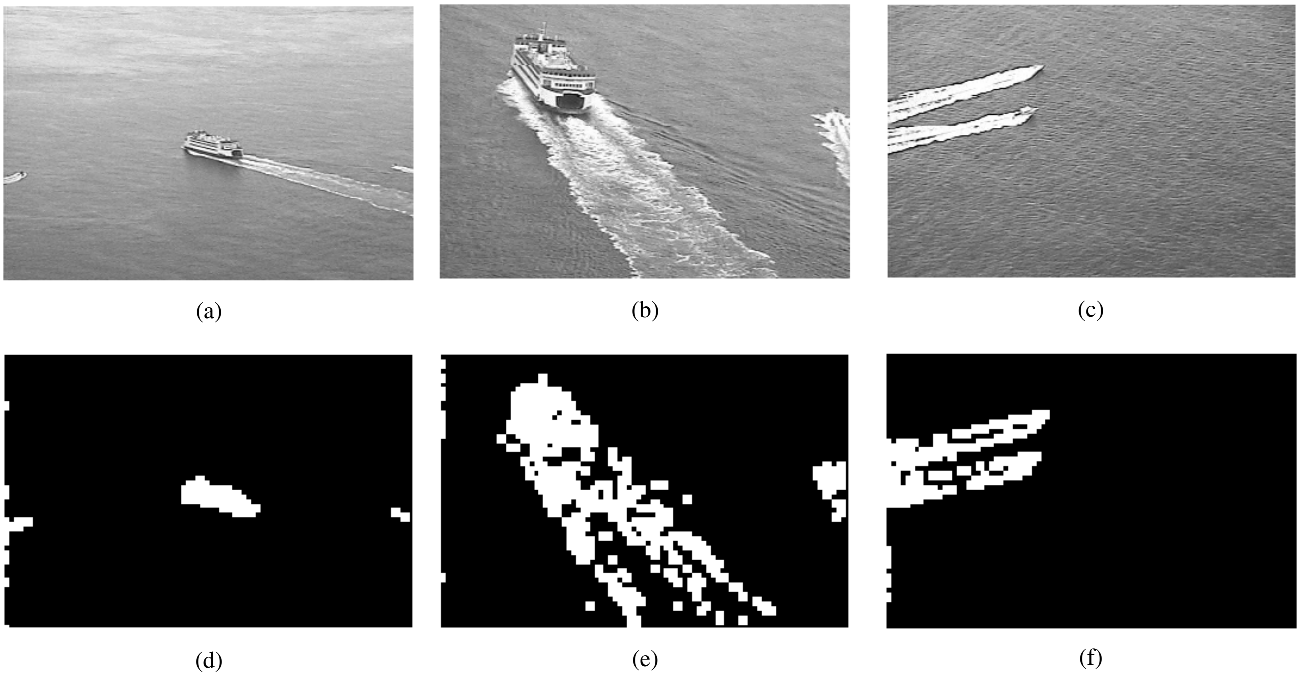

Fig. 4: Results obtained using DCT-QDA detection algorithm. Top row (a-c) shows the grayscale images, and the bottom row (d-f) are final masks obtained after filtering

make the assumption that data from all classes share the same distribution and the same covariance. Therefore, given a set of training samples with known class labels from all the classes, QDA estimates sample covariance for each class separately and uses the likelihood ratio test to determine the class label of a test sample with unknown label. Specifically, for a binary classification problem with labels $\{0, 1\}$, the likelihood ratio test to classify a test sample $\mathbf{x}$ is given by

$$\lambda(\mathbf{x}) = \frac{\sqrt{|\Sigma_1|}^{-1} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_1})^T \Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu_1})\right)}{\sqrt{|\Sigma_0|}^{-1} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_0})^T \Sigma_0^{-1}(\mathbf{x}-\boldsymbol{\mu_0})\right)} \quad (5)$$

and,

$$label(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad \lambda(\mathbf{x}) > t \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where Σ_i and $\mu_\mathbf{i}$ represent the covariance and mean of class i, and t is some threshold. This is an example of a supervised classification methodology and hence requires labeled training examples. In order to keep the spirit of minimal prior knowledge scenario, only four frames per video are weakly labeled (with a loose bounding box around the target) and used as the training samples while the rest of the video is set aside as the test set. Please note that only the DCT feature vectors extracted from training frames are used in the estimation of class-specific covariance matrices and means in Eq. 6

Fig. 4 shows some results obtained from using this detection method.

F. Blob Analysis and Target Tracking

The binary signal from the anomaly detection stage is then run through blob analysis algorithms to determine which detected anomalies are legitimate targets. This is done by applying signal processing routines (e.g. eroding and dilating) on the binary mask image obtained after anomaly detection. This ensures that spurious detects caused by glitter and other noise in the background are filtered out. Next, each blob is subjected to further filtration based on appropriate size and shape (in the cases when these parameters are known about the target) or position within the frame. This process removes most artifacts from the video, such as borders or stripes across the frame. Following this, a tracking-by-detection strategy is employed on the anomalies that qualify as objects of interest after blob analysis. This step not only provides our system with object tracking capability, but also helps to make the detection more robust by leveraging information from neighboring frames. For example, a persistent state for each track is kept for analysis throughout the whole video. For each frame, a Kalman filter is used to predict the current position of each track based on previous measurements. Blobs from the current frame are then matched to existing tracks based on their estimated location and image features. If a blob cannot be matched to an existing track, a new track is generated with that blob as the first point. If a track has no blobs matched to it for several frames ($\sim$30) it is considered lost and is archived. Finally, at the end of the process, each track is updated and corrected with the information added by matched blobs.

IV. Conclusion and Future Work

We have described a system developed for robust multi-scale marine object detection in full-motion video (FMV) of unknown video quality from Unmanned Aerial Vehicles (UAV). Currently, computer vision and object detection research is primarily carried out on high quality videos, the motion of one or more of the systems involved is well-understood, and typically leverages on prior knowledge of target size or appearance. Due to the nature of our data and application domain, our approach makes no assumptions about the systems involved and attempts to leverage signal processing and general anomaly detection algorithms to construct a generic marine object detection system that is robust to variances in illumination, foreground-background characteristics, video quality, target size and appearance.

Our future work will investigate the applicability and effectiveness of other features and methodologies to detect marine objects in full-motion videos recorded by UAVs. We are also testing our system on videos recorded by other platforms, such as ship-mounted surveillance cameras, that provide highly oblique perspectives.

Acknowledgments

This work is supported by the Naval Innovative Science and Engineering (NISE) program at the Space and Naval Warfare Systems Center Pacific. The authors would like to thank Josh Harguess, Eric Hallenborg, and Katie Rainey for helpful discussions.

References

[1] H. Buck, E. Sharghi, K. Bromley, C. Guilas, and T. Chheng, *Ship detection and classification from overhead imagery*, In Proc. SPIE 6696, 2007.

[2] H. Buck, E. Sharghi, C. Guilas, J. Stastny, W. Morgart, B. Schalcosky, and K. Pifko, *Enhanced ship detection from overhead imagery*, In Proc. SPIE 6945, 2008.

[3] J. Stastny, M. Hughes, D. Garcia, B. Bagnall, K. Pifko, H. Buck, and E. Sharghi, *A novel adaptive synthetic aperture radar ship detection system*, OCEANS, pp. 1-7, 2011.

[4] T. P. Breckon, S. E. Barnes, M. L. Eichner, and K. Wahren, *Autonomous real-time vehicle detection from a medium-level UAV*, In Proc. 24th International Conference on Unmanned Air Vehicle Systems, pp. 29.1-29.9, 2009.

[5] A. Gaszczak, T.P. Breckon, and J.W. Han, *Real-time people and vehicle detection from UAV imagery*, In Proc. SPIE Conference Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques, Volume 7878, No. 78780B, 2011.

[6] K. Kaaniche, B. Champion, C. Pegard, and P. Vasseur, *A vision algorithm for dynamic detection of moving vehicles with a UAV*, In Proc. Int. Conf. on Robotics and Automation, pp. 1878-1883, 2005.

[7] J. Xiao, C. Yang, F. Han, and H. Cheng, *Vehicle and person tracking in aerial videos*, In Proc. Int. Workshop on Multimodal Technologies for Perception of Humans, pp. 203-214, Berlin, Heidelberg, 2008.

[8] P. Malinoski, *Object detection and recognition for UAV*, Thesis, 2012.

[9] S. Fefilatyev, *Detection of marine vehicles in images and video of open sea*, Thesis, 2008.

[10] S. Zhang, Z. Qi, D. Zhang, *Ship tracking using background subtraction and inter-frame correlation*, 2nd International Congress on Image and Signal Processing, 2009. CISP '09., pp.1-4, 2009.

[11] X. Yu, I. S. Reed and A. D. Stocker *Comparative performance analysis of adaptive multispectral detectors*, IEEE Transactions on Signal Processing, 41(8), pp. 2639-2656, 1993.

[12] K. Rao and P. Yip *Discrete Cosine Transform: Algorithms, Advantages, Applications*, Academic Press, ISBN 0-12-580203-X, 1990.

[13] T. Hastie, R. Tibshirani and J. Friedman *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer Series in Statistics. Springer-Verlag, New York, Second Edition, 2009.

[14] M. Siam, M. ElHelw *Robust autonomous visual detection and tracking of moving targets in UAV imagery*, IEEE 11th International Conference on Signal Processing (ICSP), vol.2, pp. 1060-1066, Oct. 2012