Multiresolution spot detection by means of entropy thresholding

Giuseppe Boccignone

Dipartimento di Ingegneria dell'Informazione e Ingegneria Elettrica and Istituto Nazionale di Fisica della Materia, Università di Salerno via Ponte Melillo 1, Fisciano, Italy

Angelo Chianese and Antonio Picariello

Dipartimento di Informatica e Sistemistica, Università di Napoli "Federico II," via Claudio 21, Napoli, Italy

Received May 6, 1999; revised manuscript received April 3, 2000; accepted April 7, 2000

Many imaging applications deal with the detection of small targets or spots embedded within an inhomogeneous background. We present a method that accomplishes a multiresolution detection on the wavelet-transformed image. The targets are separated from the background by the exploitation of Renyi's information, which is evaluated at the different decomposition levels of the wavelet transform. The scale-dependent candidate detections are successively combined by means of majority voting for final detection. Connections with results provided in different fields such as multifractal analysis, generalized information measures in scale-space, and cross-entropy analysis in fine-to-coarse transformations are discussed. Detection performance is investigated through an example from medical image analysis.
[S0740-3232(00)02107-4]

OCIS codes: 100.0100, 100.2000, 100.2960, 100.7410.

1. INTRODUCTION

In this paper we study the detection of small targets embedded within an inhomogeneous background. The problem is of interest for different applications, such as the detection of tumors and other irregularities in medical images and target detection in infrared sensor data.[1–4] Figure 1 provides an example from digital mammography. The figure shows a rectangular area extracted from a mammogram, which consists of two kinds of regions: microcalcifications, representing the targets to be detected, and background tissue. The microcalcifications are shown as brighter spots, whose actual average size is ~0.3 mm in diameter, hidden within "noise" originated from both the acquisition process and the broad variations of breast tissue structure. The very small extension of these targets requires a mammogram to be spatially sampled at least at a pixel size of 0.1 mm, whereas the detection accuracy is independent of digitization depth in the range from 12 to 9 bits per pixel.[5] Such an example is a challenging one for a machine and extremely so for a human observer. It has been claimed that, as normally viewed, images of this kind display only ~3% of the information they detect.[6] In practice, beyond the diagnostic quality of the images, visual assessment depends on factors relating to the diagnosticians, such as their level of training, use of search versus nonsearch protocols, their premature termination of the search, and use of conservative decision criteria.[7] In this framework, computer-aided detection is reputed to be a promising line of investigation.[7]

Note that, in principle, the detection task should be distinguished from the segmentation task, the latter accounting for target localization accomplished together with exact shape reconstruction. Here we focus on the spatial localization process; the terms detection and segmentation will be used interchangeably. To face the problem, one could assume prior, context-based knowledge to be available and then compute Bayesian estimates, which are optimal under the specific "true" model, for the unknown signal. For instance, in mammography a method has been proposed on the basis of Markov random fields,[8] which allows prior knowledge of spatial relations among objects to be explicitly modeled but at a price of high computational complexity. Actually, a complete model is seldom, if not at all, available. All of this makes particularly appealing methods, such as the one described here, carried out under the assumption of no context-based knowledge available *a priori*.[2,3]

The essence of our approach is to decompose the image on a wavelet basis and use the thresholding of the wavelet coefficients to obtain different scale-dependent estimates of the objects of interest. To this end, a wavelet representation of the original image is obtained by means of a weighted, undecimated discrete wavelet transform. Then, at each scale, candidate targets are detected by a thresholding procedure; what is novel here is that our approach automatically determines the threshold function by relying on an information-theoretic tool, namely, Renyi's information.[9] Eventually, the candidate targets collected at the different scales are suitably combined to obtain the final detection. In this way, images with varying characteristics can be successfully segmented with one procedure and without the need for *a priori* fixed thresholds. Meanwhile, image reconstruction by means of an inverse wavelet transform is avoided, thus relaxing constraints on the choice of the wavelet basis.[10]

Fig. 1. Example of an image (a mammogram) containing small targets, microcalcifications, embedded within an inhomogeneous background. A microcalcification is a tiny calcium deposit that has accumulated in the breast tissue, and it appears as a small bright spot in the mammogram.

It is worth noting that wavelet thresholding, as discussed by Donoho and Johnstone and subsequent works in the wavelet literature,[11,12] has actually been employed for denoising rather than true detection. Denoising methods work as follows: perform the wavelet transform of the observed data; apply simple thresholding nonlinearities to empirical wavelets coefficients; and compute signal estimates by taking the inverse wavelet transform of the thresholded coefficients. However, many of the procedures derived to date are sensitive to noise distributions whose tails are heavier than the Gaussian distribution, as in the case of images like mammograms. In addition, a reconstruction step is required.[12]

As concerns the idea of specifically exploiting wavelet coefficient thresholding for object localization, a method has been proposed by Strickland and Han.[4] Their motivation for using wavelets is that such bases can act as an optimum detector or matched filter. Since the matched filter technique requires specific knowledge of the target signal, they introduce the simplifying assumptions that the spots to be detected are truly Gaussian objects and the noise is one of the Markov types. By choosing the Laplacian of Gaussian wavelets, the wavelet transform performs as a multiscale matched filter and, at each scale, the output of the different subbands can be thresholded to produce a detect/no detect result. It can be noted that our work shares some aspects with Strickland and Han's[4]; therefore, their results are a suitable reference for comparison. In contrast, the overall rationale is quite different; most important, in Strickland and Han's approach no general estimation problem is addressed (in the denoising sense). The main drawback of such an approach is that the detection threshold is experimentally chosen as a fixed percentile of the histogram of each channel. Further, in many applications, targets will occur at unpredictable scales and their shape will be non-Gaussian. In this case, one has to take into account the problem of using the features extracted at the different scales as a suitable input to a classifier.

We begin in Section 2 with a formal statement of the problem and a discussion in depth of the assumptions that motivate our approach; in addition, for the sake of completeness, we give a short review of some basic concepts of wavelet analysis that are used throughout the paper. In Section 3 we present the method. In Section 4 we report outcomes of experimental work performed on a standard mammogram data set. In Section 5 we discuss results so far achieved and provide a physical interpretation of the approach, which highlights some links to multifractal analysis, information measures in scale-space,[13] and cross-entropy evaluation in fine-to-coarse transformations.[14] We conclude in Section 6 with a review and final comments.

2. BACKGROUND

Define the image f as a random field from D to a subset of the real numbers $\mathbf{R}$, where the image domain $D \subseteq \mathbf{Z}^2$, $\mathbf{Z}$ being the set of integers, is a lattice of N sites; let $\mathcal{F} = \{F_p\}_{p \in D}$ be any family of P random variables indexed by $p \in D$. A possible sample f of F_p is denoted by $f = (f_{p_1},\ldots, f_{pP})$, and it is called a configuration of the field. The value of the field observations represent pixel intensity; namely, $f_p = f(n, m)$, where $(n, m) \in \mathbf{Z}^2$ specifies the coordinates of site p, provided that $P = N \times M$, if $1 \leqslant n \leqslant N$ and $1 \leqslant m \leqslant M$. Namely, N and M are the horizontal and vertical dimensions of the image.

Consider the image as the collection of two sets of pixels D^τ and D^β related to the targets and to the background, respectively; the sets are disjoint and cover all the domain D, that is, $D^\tau \cap D^\beta = \varnothing$ and $D = D^\tau \cup D^\beta$. Then a segmentation is a partition of D through the labeling of the field f as $M(f) = [M^\tau(f_{p_k}), M^\beta(f_{p_j})]_{p_k \in D^\tau,\, p_j \in D^\beta}$, such that $M^\tau \cap M^\beta = \varnothing$, where $M^\tau = \cup_{k \in D^\tau} M^\tau(f_{p_k})$ and $M^\beta = \cup_{j \in D^\beta} M^\beta(f_{p_j})$. Such a segmentation can be reformulated as an estimation problem of the true signal f_τ, given a background noise f_β and observed data f:

$$f(n, m) = f_\tau(n, m) + f_\beta(n, m). \tag{1}$$

Equation (1) provides a simplified model in which the noisy term is due to the joint contribution of actual noise and background structures. For instance, in a digitized mammogram, the actual noise component should account for different contributions: the blur due to patient movement; a finite focal spot; intensifying screen glare; scattered radiation, which acts as a contrast reducer; and dust and dirt, which lies between the film and the intensifying screen and attenuates the light before exposing the film. Further, the distribution of such noise is complex, since it is blended with tissue structures, and its variation in different regions depends on the predominant type of tissue. For example, the noise level is considerably higher in bright regions of the mammogram representing glandular tissue. Likewise, in infrared images, f_β appears as a mix of correlated and uncorrelated noise components; notwithstanding, background structures could be predominant in certain regions. In the literature, such a distinction is often taken into account by distinguishing between clutter and noise: The components of f_β that are stochastic random processes are called noise, and the deterministic components that are not part of the signal f_τ are identified as clutter.[2] In the most general case, the background signal is treated as a nonideal background and modeled as a cloud signal of fractal nature.[2]

Clearly, if we do not consider any context-based information (for instance, the geometric characteristics of the clutter), the amount of information at hand is a function of the signal-to-noise ratio.[15] Unfortunately, for complex, nonstationary images, the classic definition of an average signal-to-noise ratio is of moderate interest. More attractive is the local signal-to-noise ratio (LSNR) defined as[3]

$$\mathrm{LSNR(dB)} = 10 \log_{10} \sum_{k,\, j \in \partial_{(n,m)}} \times [f_\tau(n - k, m - j) - \mu_\beta]^2/\sigma_\beta^2,$$

where σ_β^2 and μ_β are the variance and the mean of the background, respectively, measured in the window $\partial_{(n,m)}$ around the pixel of interest (n, m). Provided that f_τ is known, when the LSNR is gauged at different parts of a mammogram, e.g., Fig. 1, or on simulated cloud data,[3] a high variability is counted from point to point (e.g., $-3\ \mathrm{dB} \div 16\ \mathrm{dB}$). However, for the problem we address here, where f_τ and f_β are initially unknown, a more appropriate measure is the LSNR defined in terms of the local contrast ratio (LCR)[8]; namely,

$$\mathrm{LCR(dB)} = 10 \log_{10} N_\partial \times \left[f(n, m)/\sum_{k,\, j \in \partial_{(n,m)}} f(n - k, m - j) \right],$$

N_∂ being the size of the window $\partial_{(n,m)}$. It is important to note that in both LSNR and LCR one has to choose the size of the window $\partial_{(n,m)}$ and that different choices may lead to different measures. This makes clear one point: For the problem we are facing, a signal-to-noise ratio definition does not make sense until one defines the scale (window size) of the target with respect to the scale of the background features.

With this rationale supporting us, we often find it advantageous to work in a transform domain rather than attempting to separate signal from noise in the spatial (pixel) domain. The general motivation for transform domain processing is that an appropriate image transform can concentrate the signal energy into a small number of coefficients. Two standard choices for the image transform are the Fourier transform (FT) and the wavelet transform (WT). If a FT is adopted, one should cope with the fact that in the spatial signal the harmonic frequency components are present but hidden, whereas in the spatial-frequency spectrum it is the spatial information that is hidden. In contrast, the targets should be preserved under a transform that can localize the signal characteristics in the original and the transformed domain. A suitable transform that satisfies these requirements is the WT. The WT uses basis functions that can dilate in scale and translate in position according to signal characteristics.[10] These basis functions have a remarkable aptitude for localizing the signal in both spatial and spatial-frequency domains, in the sense of the classical uncertainty principle.[16] Moreover, empirical evidence has shown that wavelet bases provide more efficient representations of real-world data than either spatial or frequency domain representations.[17] Because wavelets are able to represent complicated signal structure concisely, filtering techniques based in the wavelet domain are often much better at separating signals from noise than classical approaches based in the spatial or frequency domain alone, especially when the signal and noise have overlapping spectral content. Intuitively, in the WT domain a few coefficients will have high LCR and can be kept, whereas a large number of coefficients, which can be discarded, will exhibit low LCR. In this sense, wavelets implement the idea that the signal-to-noise ratio should be tuned on the scale of the target with respect to the scale of the background features.

We recall that a one-dimensional (1-D) wavelet $\psi(x)$ is a function of zero average, $\int_{-\infty}^{+\infty} \psi(x) \mathrm{d}x = 0$, which is dilated with a scale parameter s and translated by u as $\psi_{s,u}(x) = (1/\sqrt{s})\, \psi[(x - u)/s]$. The functions $\{\psi_{s,u}\}$ form a family of basis functions, and the WT of a signal f is the decomposition of f on this basis computed as the inner product

$$W_s f(u) = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} f(x) \psi^* \left(\frac{x - u}{s} \right) \mathrm{d}x, \qquad (2)$$

where the $1/\sqrt{s}$ factor is used to conserve the norm. In the following we will consider real bases; thus $\psi^* = \psi$. Equation (2) indicates that $W_s f(u)$ is obtained when the convolution of the signal with the chosen wavelet is performed. Since ψ has by definition a zero average, the convolution measures the variation of f in a neighborhood of u, whose size is proportional to scale s. Intuitively, this transformation can be interpreted as a mathematical microscope whose position and magnification are u and $1/s$, respectively, and whose optical characteristic is given by the choice of the specific wavelet ψ. Two issues are important for this choice: compact support and regularity. Compact support improves the spatial resolution of wavelets. Wavelets are compactly supported if they have finite support with a maximum number of vanishing moments, for they support width. A wavelet has n vanishing moments if $\int x^k \psi(x) \mathrm{d}x = 0$, for $0 \leqslant k < n$. Regularity relates to differentiability: Since differentiation in the Fourier domain amounts to a multiplication by $(j\omega)$, the existence of derivatives is related to sufficient decay of the Fourier spectrum.

The basic idea of wavelet analysis is to represent any arbitrary function f as a superposition (integral) of wavelets. In practice, f is written as a discrete superposition (sum). To this end, one can define a discrete scaling $s = s_0^l$, based on a dilatation step $s_0 > 1$, and let the translational parameter u vary with s and the translational step u_0; that is $u = n u_0 s_0^l$, where $l,\ n \in \mathbf{Z}$. Then the family of basis functions $\psi_{s,u}$ can be written as $(1/s_0^{l/2})\, \psi[(x - n u_0 s_0^l)/s_0^l]$. The WT can be more or less redundant, depending on the type of sampling (s_0, u_0 values). In the case of critical sampling $u_0 = 1$ and $s_0 = 2$ (i.e., the scale s is sampled along a dyadic sequence $\{2^l\}_{l \in \mathbf{Z}}$), the wavelet family can be chosen as the real orthonormal basis $\{\psi_{l,n}(x) = 1/2^{l/2}\, \psi(2^{-l}x - n)\}$. Consequently, any finite energy signal can be decomposed over this basis as: $f = \sum_{l=-\infty}^{+\infty} \sum_n \langle f,\ \psi_{l,n} \rangle \psi_{l,n}$.

The algorithm proposed by Mallat[18] relates an orthogonal WT with multiresolution analysis. In a nutshell, the mother wavelet is constructed with a low-pass scaling function $\phi(x)$ that satisfies the two-scale equations $\phi(x) = \sqrt{2}\Sigma_n h(n)\phi(2x - n)$, $\psi(x) = \sqrt{2}\Sigma_u g(n)\phi(2x - n)$, where $g(k) = (-1)^k h(1 - k)$. The coefficients $h(k)$ and $g(k)$ have to meet several conditions for the set of basis functions to be unique and orthonormal and have a certain degree of regularity. Then the 1-D signal can be represented as

$$f(x) = \sum_n c_{L,n} 2^{-L/2}\phi(2^{-L}x - n) + \sum_{j=-\infty}^{L} 2^{-l/2}\sum_n w_{l,n}\psi_{l,n}(x), \tag{3}$$

where $c_{l,n} = 2^{-l/2}\int f(x)\phi(2^{-L}x - n)\mathrm{d}x$ are the scaling coefficients and $w_{l,n} = 2^{-l/2}\int f(x)\psi(2^{-L}x - n)\mathrm{d}x$ are the wavelet coefficients. Here L denotes the maximum, coarse scale. For a dyadic decomposition of a N-samples signal, $L = \log_2 N$.

Filtering operations can be performed by use of the $h(k)$ and $g(k)$ coefficients as the impulse responses that correspond to low- and high-pass operations, respectively. The scaling and wavelet coefficients can be recursed as $c_{l+1,u} = \Sigma_n h(2u - n)c_{l,n}$ and $w_{l+1,u} = \Sigma_n g(2u - n)c_{l,n}$. Noting that these recursive equations define the convolutions $c_{l+1,u} = c_{l,n} * (2u)$ and $w_{l+1,u} = c_{l,n} * g(2u)$, respectively, such computations can be succinctly organized into a discrete-time filter bank comprising low-pass and high-pass filters and decimators.

On the other hand, if one is tied to specific constraints, e.g., the linear phase must be preserved to achieve wavelet symmetry, then the orthonormality requirements must be relaxed by the employment of biorthogonal (quasi-orthogonal) bases. This means that reconstruction wavelets $\tilde{\psi}_{l,n}$ must be used, which may be different from decomposition ones; i.e., $f = \Sigma_{l=-\infty}^{+\infty}\Sigma_n\langle f, \psi_{l,n}\rangle\tilde{\psi}_{l,n}$. Consequently, a scaling function $\tilde{\phi}$ together with synthesis filters $\tilde{h}$ and $\tilde{g}$ should be appropriately selected.

To deal with images, we need a two-dimensional (2-D) WT. The 2-D wavelet bases are expressed as the product of two 1-D wavelet bases along the horizontal and vertical directions. More precisely, from the 1-D scaling function ϕ and the corresponding wavelet ψ, we construct the set $\{\psi^k(x, y)\}_{1\leqslant k\leqslant 3}$, consisting of the three functions $\psi^1 = \phi(x)\psi(y)$, $\psi^2 = \psi(x)\psi(y)$, and $\psi^3 = \psi(y)\psi(x)$, from which by scaling and dilation we obtain the basis $\{\psi^k_{l,n,m}(x, y) = 1/2^l\,\psi^k(2^{-l}x - n, 2^{-l}y - m)\}$ for n, m running through $\mathbf{Z}^2$. The 2-D scaling function is computed as $\phi(x, y) = \phi(x)\phi(y)$.

The 2-D WT is then defined by the inner product $W_{k,s}f = \langle f, \psi^k_{l,n,m}\rangle$, where the scale index l is not greater than $L = \log[\min(N, M)]$. Thus a finite energy image can be decomposed by use of separable 2-D wavelet bases according to the fundamental identity

$$f = \sum_l \frac{1}{2^l}\left(\sum_k \sum_{n,m}\langle f, \psi^k_{l,n,m}\rangle\psi^k_{l,n,m}\right). \tag{4}$$

Just like in the 1-D case, a precise identity as specified by Eq. (3) can be stated. In other terms, the 2-D WT kernel is a separable kernel function. Typically, 2-D wavelet bases are implemented with a 2-D filter bank in which the application of the h and g filters is alternated on rows and columns of the image $f(n, m)$.[12] The corresponding filter coefficients are expressed as $h_{\mathrm{LL}}(n, m) = h(n)h(m)$, $h_{\mathrm{HL}}(n, m) = g(n)h(m)$, $h_{\mathrm{LH}}(n, m) = h(n)g(m)$, and $h_{\mathrm{HH}}(n, m) = g(n)g(m)$, where the first and second subscripts represent the low-pass and high-pass filtering effects in the x and y direction, respectively.

Formally, denote, at all scales $2^l \geqslant 1$, $w^k_l(n, m) = \langle f, \psi^k_{l,n,m}\rangle$ and $c_l(n, m) = \langle f, \phi_{l,n,m}\rangle$ the wavelet coefficients and the scaling coefficients, respectively, where $c_0(n, m) = \langle f, \phi_{0,n,m}\rangle$ is by definition the original image $f(n, m)$. With separability of the wavelet basis recursion equation for coefficients, it has been shown that the 2-D WT can be computed by iterating for $0 \leqslant l < L$ the following: $c_{l+1}(n, m) = c_l * h_{\mathrm{LL}}(2n, 2m)$, $w^1_{l+1}(n, m) = c_l * h_{\mathrm{LH}}(2n, 2m)$, $w^2_{l+1}(n, m) = c_l * h_{\mathrm{HL}}(2n, 2m)$, and $w^3_{l+1}(n, m) = c_l * h_{\mathrm{HH}}(2n, 2m)$. Here c_0 denotes the original image f.

Practically, the rows of c_l are first convolved with h and g and subsampled by two. The columns of these two output images are then convolved with h and g, respectively, and subsampled, which gives the subsampled images c_l, w^1_l, w^2_l, and w^3_l. The resulting representation of the image is eventually composed of the $3L + 1$ subimages or coefficient planes $(c_L, \{w^k_l\}_{1\leqslant l\leqslant L, 1\leqslant k\leqslant 3})$. The plane c_l (low-frequency subband) represents the smoothed image. For $k = 1, 2, 3$, the detail subbands contain horizontal (LH), vertical (HL), and diagonal (HH) directional information, respectively.

This algorithm, known as the fast 2-D WT, might not be convenient for pattern-detection purposes: On the one hand, it is not shift invariant; on the other hand, image subbands result uncorrelated at the different scales. To overcome such drawbacks, an undecimated discrete WT has been proposed.[19] Such scheme is characterized by subbands obtained by alternating a low-pass filter tapped by a bandpass filter (as in the above algorithm) but inserting $2^l - 1$ zeros between each sample of the filter in place of decimation. In this case, the subbands remain at full size, and this spatial redundancy, which may turn in a loss of computational efficiency of subsequent processing, is clearly counterbalanced by a higher suitability of the representation for pattern-detection aims. Here we will extend the above scheme to a weighted, undecimated discrete WT.

As a first step, an undecimated discrete transform is applied to the image f. Note that in our case the detail planes $w^k_l = \{w^k_l(n, m)\}$ are obtained from the image equation (1), which by linearity gives

$$w^k_l(n, m) = w^k_l(n, m)_\tau + w^k_l(n, m)_\beta. \tag{5}$$

Each plane has the same dimensions of the original image, that is $N \times M$. The fact we are not concerned with exact signal reconstruction reduces the transform constraints. First, we can adopt, along the analysis phase, either an orthogonal or a biorthogonal basis, without being concerned with the properties of the synthesis basis.

Second, we can further compress and specifically tune the representation obtained, by suitably combining the undecimated detail subbands. Namely, at each level l, a representation plane $w_l = \{w_l(n, m)\}$ is built, where each coefficient is computed as the weighted linear combination

$$w_l(n, m) = \sum_{k=1}^{3} \gamma_k |w_l^k(n, m)|. \tag{6}$$

Consequently, we achieve a multiresolution representation, at which, at each scale l, we have one coefficient plane $\{w_l(n, m)\}$, each coefficient modeled as

$$w_l(n, m) = w_l(n, m)_\tau + w_l(n, m)_\beta. \tag{7}$$

In the sequel we will not consider the low-frequency band c_l, since the searched information mostly resides in the detail subbands.

3. THEORY

If the signal as described by Eq. (1) is a realization of a non-Gaussian process, one can use a nonlinear estimation to recover f_τ from noisy measurements. First, decompose the observed signal f [Eq. (1)], in an orthogonal basis $\mathcal{B} = \{b_v\}_{0 \leqslant v \leqslant V}$:

$$\langle f, b_v \rangle = \langle f_\tau, b_v \rangle + \langle f_\beta, b_v \rangle. \tag{8}$$

Then the estimator $\tilde{f}_\tau$ can be computed as $\tilde{f}_\tau = \Sigma_{v=0}^{V-1} \langle f, b_v \rangle t(v) b_v$, where $t(v)$ is a set of constant coefficients that in the ideal case are chosen as $t(v) = 1$, if $|\langle f_\tau, b_v \rangle|^2 \geqslant \sigma^2$, and $t(v) = 0$ otherwise, σ^2 being the noise variance. The estimator performance is measured by the mean-square error

$$\epsilon_s = E(\|f_\tau - \tilde{f}_\tau\|^2) = \sum \min(|\langle f_\tau, b_v \rangle|^2, \sigma^2). \tag{9}$$

Clearly, nonlinear estimators are signal dependent and require, in principle, a precise modeling of the signal process. Unfortunately, ideal coefficient selection cannot be implemented because values $\langle f_\tau, b_v \rangle$ are not known in advance. However, it has been shown that nonlinear thresholding in orthogonal basis performs as a nearly optimal estimator.[11]

Assume that the orthogonal basis $\mathcal{B}$ has been chosen to concentrate signal energy over a few coefficients having a high amplitude above the noise; this can be done either experimentally, along a training phase, or, if possible, analytically.[17] A thresholding estimator can be defined as[11]

$$\tilde{f}_\tau = \sum_{v=0}^{V-1} t(\langle f, b_v \rangle) b_v, \tag{10}$$

where $t(\langle f, b_v \rangle) = \langle f, b_v \rangle$ if $|\langle f, b_v \rangle| > t_{\text{opt}}$ and $t(\langle f, b_v \rangle) = 0$ otherwise, t_{opt} being an optimal threshold. The performance of such an estimator has been proved to be close to the ideal coefficient selection.[11] In particular, the mean-square error ϵ,

$$\epsilon = \sum E[|\langle f, b_v \rangle - t(\langle f, b_v \rangle)|^2], \tag{11}$$

is of the same order of the ideal coefficient selection ϵ_s [Eq. (9)].

Notably, Donoho and Johnstone[11] have shown that such a performance is related to the rate of decay of the sorted decomposition coefficients. Most interesting, if the basis $\mathcal{B}$ is chosen as an orthogonal wavelet basis, the coefficient decay is related to regularity (structure) of the signal and precisely to its Lipschitz regularity (cf. Section 5). When the observed signal f is decomposed in a discrete wavelet basis, thresholding as given in Eq. (10) becomes

$$\tilde{f}_\tau = \sum_l \frac{1}{2^l} \left[\sum_k \sum_{n,m} t\{\langle f, \psi_{l,n,m}^k \rangle\} \psi_{l,n,m}^k \right]. \tag{12}$$

Signal estimation performed according to Eq. (12) requires an orthogonal basis and an inverse WT. To overcome such constraint, it is possible to estimate the signal in the transform domain directly, with coefficients defined by Eq. (7), as follows.

Denote $\tau_l(x)$ and $\beta_l(x)$, $0 \leqslant \tau_l(x)$, $\beta_l(x) \leqslant 1$, the distributions (histograms) of the wavelet coefficients $x = w_l(n, m)$ relative to objects and background, respectively. The distributions $\tau_l(x)$ and $\beta_l(x)$ are defined on the supports Ω_l^τ and Ω_l^β, respectively. Both these supports depend on the parameter t_l, namely, $\Omega_l^\tau(t_l) = \{x : x_{l\,\min} \leqslant x \leqslant t_l\}$ and $\Omega_l^\beta(t_l) = \{x : t_l < x \leqslant x_{l\,\max}\}$, where $x_{l\,\min} = \min\{w_l(n, m)\}$ and $x_{l\,\max} = \max\{w_l(n, m)\}$. Thus we suppose that $\tau_l(x) \rightarrow 0$ on $\Omega_l^\beta(t_l)$ and that $\beta_l(x) \rightarrow 0$ on $\Omega_l^\tau(t_l)$.

To detect the objects of interest, one should determine t_l such that the distance between the two distributions be maximized. We say that such distance represents the level l information (precisely 2^l), and we define it as the $\mathbf{l}^2$ distance $\Sigma_{x \in \Omega_l} [\tau_l(x) - \beta_l(x)]^2$ over the support $\Omega_l = \Omega_l^\tau(t_l) \cup \Omega_l^\beta(t_l)$. Clearly, since

$$\sum_{x \in \Omega_l} [\tau_l(x) - \beta_l(x)]^2 = \sum_{x \in \Omega_l^\tau} \tau_l(x)^2 + \sum_{x \in \Omega_l^\beta} \beta_l(x)^2 - 2 \sum_{x \in \Omega_l} \tau_l(x) \beta_l(x), \tag{13}$$

the term at the left-hand side of Eq. (13) is maximized if $\Sigma_{x \in \Omega_l} \tau_l(x) \beta_l(x)$ is minimum. In fact, for $\Sigma_{x \in \Omega_l} \tau_l(x) \beta_l(x) = 0$ the two distributions can be considered orthogonal (we exactly distinguish the spots from the background). If the left-hand-side term of Eq. (13) denotes "information," the last term on the right-hand side can be conceived as a sort of entropy, which destroys information while increasing. Define

$$H = \log \sum_{x \in \Omega_l} \tau_l(x) \beta_l(x) \tag{14}$$

as the entropy we want to minimize for maximizing information. From definition (14), by use of the Cauchy–Schwartz inequality $[\Sigma_{x \in \Omega_l} \tau_l(x) \beta_l(x)]^2 \leqslant \Sigma_{x \in \Omega_l} \tau_l(x)^2 \Sigma_{x \in \Omega_l} \beta_l(x)^2$ and use of logarithms, entropy H can be upper bounded as

$$2H \leqslant \log \sum_{x \in \Omega_l^\tau} \tau_l(x)^2 + \log \sum_{x \in \Omega_l^\beta} \beta_l(x)^2. \quad (15)$$

Let

$$I_{\Omega^p}(r) = \frac{1}{1-r} \log\left[\sum_{x \in \Omega^p} p(x)^r\right] \quad (16)$$

be the rth order Renyi's information[9] of the distribution $p(x)$ over the support Ω^p. Then, from inequality (15), it follows that $H \leqslant \frac{1}{2}\log \Sigma_{x\in\Omega_l^\tau}\tau_l(x)^2 + \frac{1}{2}\log \Sigma_{x\in\Omega_l^\beta}\beta_l(x)^2$, i.e.,

$$H < -[I_{\Omega_l^\tau}(2) + I_{\Omega_l^\beta}(2)]. \quad (17)$$

Thus we have shown that the minimization of entropy H can be achieved by maximizing the second-order Renyi's information of $\tau_l(x)$ and $\beta_l(x)$ with respect to the parameter (threshold) t_l:

$$t_l = \arg\max \lim_{r\to 2}\left\{\frac{1}{1-r}\log\left[\sum_{x\in\Omega_l^\tau}\tau_l(x)^r\right] + \frac{1}{1-r}\log\left[\sum_{x\in\Omega_l^\beta}\beta_l(x)^r\right]\right\}. \quad (18)$$

It is worth noting that Renyi's information is also known as generalized information or generalized entropy. In fact, if in Eq. (16) we take the limit as $r \to 1$, we obtain the Shannon entropy function $I_{\Omega^p}(1) = \Sigma_{x\in\Omega^p} p(x)\log p(x)$, whereas as $r \to 0$, we obtain the logarithm of the volume of the support set $I_{\Omega^p}(0) = \log(\mu)\{x{:}p(x) > 0\}$, that is, the measure of the support set of the density p.[20] These properties will be useful in providing a general discussion of the physical meaning of Renyi's information (cf. Section 5).

An interesting question concerns the characterization of the thresholding performance. In the following we show how, by using the threshold as defined in Eq. (18), we obtain an estimator whose performance is comparable with that proposed by Donoho and Johnstone[11] and which is related to the objects of interest at a given scale.

Define the variance of the wavelet coefficients $x = w_l(n, m)$ as $\sigma^2 = E(x^2) - E(x)^2$. The variance of the background can be bounded as $\sigma^2 < E(x^2) = \Sigma_{x\in\Omega_l} x^2\beta_l(x) < \Sigma_{x\in\Omega_l} x^2 \Sigma_{\in\Omega_l}\beta_l(x)$. Namely, the following holds,

$$\sigma^2 < c\sum_{x\in\Omega_l}\beta_l(x), \quad (19)$$

where $c = \Sigma_{x\in\Omega_l} x^2$.

It has been previously shown that the maximization of the left-hand side of Eq. (13) is equivalent to finding the minimum of $\Sigma_{x\in\Omega_l}\tau_l(x)\Sigma_{x\in\Omega_l}\beta_l(x)$ with respect to a given t_l. From inequality (19) it follows that $\min_t \sigma^2(1/c)\,\Sigma_{x\in\Omega_l}\tau_l(x) > \sigma \min_t[(1/c) \times \Sigma_{x\in\Omega_l}\tau_l(x)]^{1/2}$. Thus the computation of t_l, performed by maximizing the right-hand side of Eq. (18), is equivalent to determine t_l by minimizing $[(1/c)\,\Sigma_{x\in\Omega_l} t_l(x)]^{1/2}$. Hence

$$t_l = \sigma \arg\min_t \left[\frac{1}{c}\sum_{x\in\Omega_l}\tau_l(x)\right]^{1/2}. \quad (20)$$

Donoho and Johnstone[11] have demonstrated that a threshold $t = K\sigma$ produces an average approximation error ϵ that remains within a K factor of the ideal coefficient selection ϵ_s, as defined in Eq. (9). They also found that $K = 2\ln P$, P being the number of elements considered in the image. In other terms, K depends on the P samples of the whole image. By taking into account Eq. (20), we can set $\bar{K} = \arg\min_t[(1/c)\,\Sigma_{x\in\Omega_l}\tau_l(x)]^{1/2}$. This shows that, in our case, the choice of $\bar{K}$ is related to the target distribution at scale l.

With t_l, at each scale the detection is performed by labeling each site (n, m) as $M_l(n, m) = 1$ if $w_l(n, m) > t_l$ and $M_l(n, m) = 0$ otherwise. Thus a resolution-dependent labeling is performed by constructing a saliency map as a binary map $M_l = \{M_l(n, m)\}$. Each connected set of nonzero locations of M_l constitutes an hypothesis target detected in the image, and all spots will be included in $M_l^\tau = \{M_l(n, m)|M_l(n, m) = 1\}$, while $M_l^\beta = \{M_l(n, m)|M_l(n, m) = 0\}$ represents the background.

Eventually, we need to work out a strategy to combine the different M_l in the final labeling or detection. To this end, we can treat the M_l maps as a feature set for input to a classifier.

Formally, assume that by using the saliency maps we are able to construct a vector of measurements, say $\mathbf{x} = \mathbf{x}(n, m) = x_1(n, m),\ldots, x_L(n, m)$, at each site of the decomposition. We then assume that we have L classifiers, one for each decomposition level. According to Bayesian theory, we should assign the label $\lambda = M^k$, $k = \tau, \beta$, provided that the *a posteriori* probability of that interpretation is maximum, i.e.,

$$P(\lambda = M^k|\mathbf{x}) = \max_{j=\tau,\beta} P(\lambda = M^j|\mathbf{x}). \quad (21)$$

The computation of the *a posteriori* probability functions would depend on the knowledge of high-ordered statistics described in terms of joint probability density functions $p(\mathbf{x}|\lambda = M^k)$, which would be difficult to infer. However, it is possible to simplify the above rule and express it in terms of decision-support computations performed by the individual classifiers. Each classifier exploits only the information conveyed by the vector $\mathbf{x}$. Since the representations used by the classifiers are distinct, it may be true for some applications that these measurements will be conditionally statistically independent.

Kittler and others[21] have shown that, when the available observational discriminatory information is highly ambiguous owing to a high level of noise, it may be appropriate, in some applications, to assume that the *a posteriori* probabilities will not deviate dramatically from the prior probabilities. Under this assumption, Eq. (21) becomes[21]

$$(1-L)P(M^k)\sum_{l=1}^{L} P(\lambda = M^k|x_l) = \max_{j=\tau,\beta}(1-L)P(M^j)\sum_{l=1}^{L} P(\lambda = M^j|x_l). \quad (22)$$

Clearly, when the *a posteriori* probabilities are hardened to produce binary-valued functions, then Eq. (22) works by combining decision outcomes rather than *a posteriori* probabilities; that is, it leads to the commonly used majority vote rule.

Finally, note that at this point we can either exploit a vector of measurements $\mathbf{x}$ at site (n, m) for which $M_l(n, m) = 1$ or directly adopt the $M_l(n, m)$ as a hard estimate of $P(\lambda = M^k|\mathbf{x})$. For the sake of simplicity, the latter form has been used throughout the experiments reported in Section 4.

4. EXPERIMENTS AND RESULTS

Experiments have been performed on the standard Nijmegen data set (information is available from the authors). This test set is formed by 40 mammograms, including both benign and malignant cases. Each mammogram is accompanied by a ground truth image produced by two expert radiologists (e.g., Fig. 2, top image).

In a preliminary stage, the detection of the single microcalcifications was considered. The aim was to tune the method's parameters for maximizing the true detections [true positive (TP)] while minimizing the false detections [false positive (FP)]. Three issues have been taken into account: (1) the choice of the wavelet bases, (2) the number of decomposition levels, and (3) the tuning of the γ_k weights in Eq. (6). Standard wavelet bases have been used: Burt–Adelson, Battle–Le Marie, B spline I, B spline II, Daubechies 4, Daubechies 6, Daubechies 8, Daubechies 10[10,17]; the best TP/FP rate, at fixed parameters, has been obtained by B-spline II wavelets. These are biorthogonal, regular wavelets of compact support.

An acceptable number of decomposition levels resulted in $L = 4$; clearly, this choice is a trade-off between the size of the object to detect and the presence of noise. As regards point (3), we initially set equal weights, i.e., $\gamma_1 = \gamma_2 = \gamma_3$. The results of these experiments show that the method detects TP's in the same regions as in the accompanying truth images and a maximum of approximately 10 FP cases are detected per image. An example is given in Fig. 2, bottom image.

A second set of experiments derived from a more subtle analysis of point (3). The γ_k value choice, $0 \leqslant \gamma_k \leqslant 1$, implies a different weighting of the information in the three subbands $k = 1, 2, 3$. A first reasonable assumption is that the strongest singularities (like background texture ones) are most likely to appear within the diagonal details band ($k = 3$) rather than within vertical ($k = 1$) and horizontal detail bands ($k = 2$). A second assumption is that subbands 1 and 2 should be equally weighted. By imposing that $\gamma_1 + \gamma_2 + \gamma_3 = 1$, with $\gamma_1 = \gamma_2$, it is possible to observe the TP/FP ratio versus the γ_3 variation in the [0,1] range [free receiver operating characteristic (FROC) curve].

For quantitative performance evaluation, the proposed method was then compared with the most recent ones, in particular Strickland and Han's.[4] The detection of microcalcification clusters was considered by counting the ratio TPC = number of TP clusters detected/number of TP clusters to be detected and the false rating FPC = number of FP clusters detected/number of images. A cluster is observed if more than two microcalcifications are localized inside a circular region of radius 0.5 cm, marked around each detected microcalcification.[8] The cluster is TP if marked as TP in the accompanying truth image; FP, otherwise. The obtained FROC is shown in Fig. 3. We obtained as best result 66% of TPC at the FPC rate of 0.7 with $\gamma_3 = 0.56$. On the same database, Strickland and Han achieve 55% of TPC at the FPC rate of 0.7. Figure 4 shows the results of detection performed on the image represented in Fig. 1, with FROC-based optimal weighting $\gamma_3 = 0.56$ and $\gamma_1 = \gamma_2 = 0.17$. Figure 4 should be compared with Fig. 2, which was obtained by equal weighting $\gamma_1 = \gamma_2 = \gamma_3$. Figures

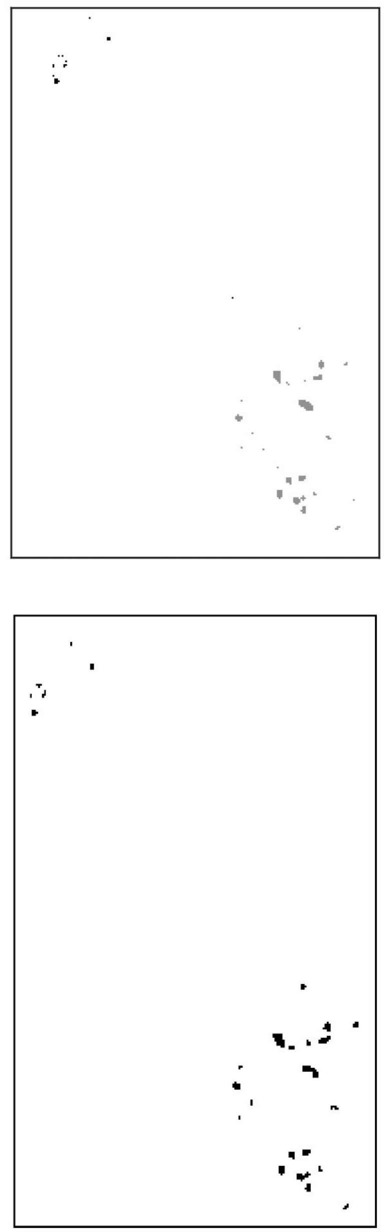

Fig. 2. Example of detection performed on the image represented in Fig. 1. Top image, the ground truth image; bottom image, the detection map.

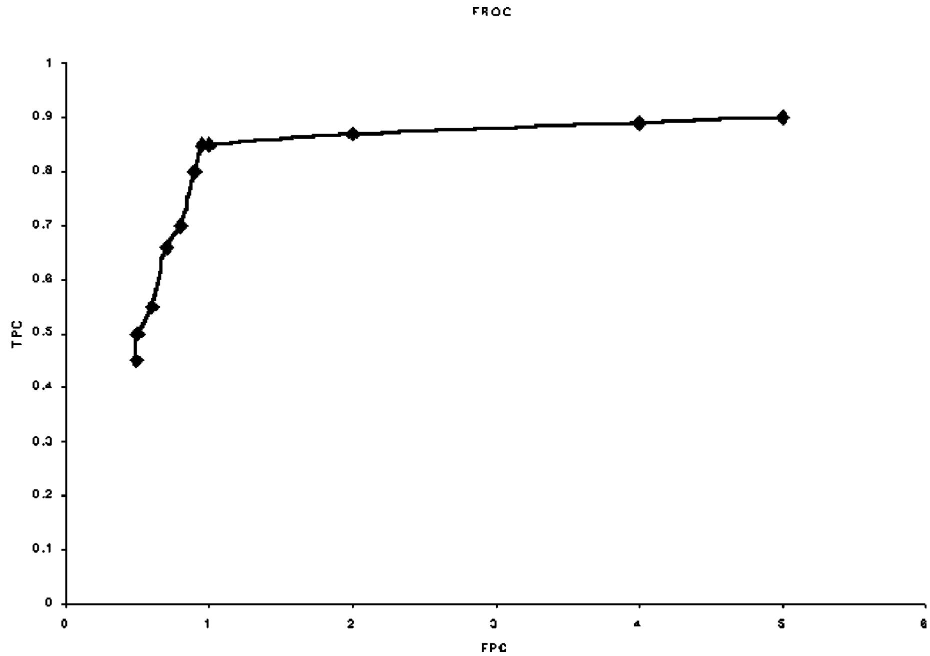

Fig. 3. Cluster detection performance done by varying the γ_3 parameter (FROC curve).

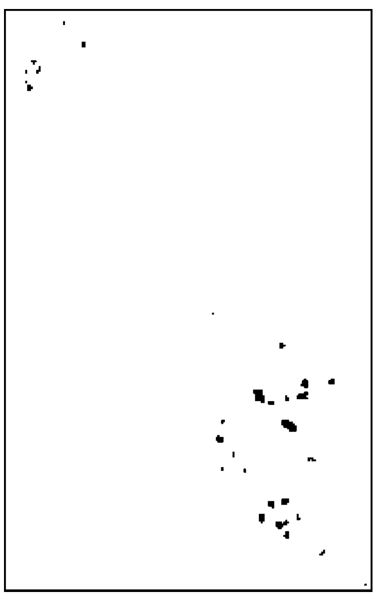

Fig. 4. Example of detection performed on the image represented in Fig. 1 but with $\gamma_3 = 0.56$ and $\gamma_1 = \gamma_2 = 0.17$.

5 and 6 display two more examples of detection produced by adopting the FROC settings.

5. DISCUSSION

An explanation of the physical meaning of the method presented in Section 3 and of the role played by Renyi information is worthwhile at this point. We have already mentioned that wavelet thresholding is equivalent to estimating the signal with a smoothing that is locally adapted to the signal regularity.[11] The wavelet theory shows that the evolution across scales of the WT depends on the local (Lipschitz) regularity of the signal. Local regularity is precisely quantified at any point in terms of the decay of the WT amplitude across scales, and such decay is related to the values of Lipschitz or Hölder exponents at that point.[12]

A function f is said to be Lipschitz α at (x_0, y_0) if there exists a constant $C > 0$ such that for all $(x, y), |f(x, y) - f(x_0, y_0)| \leqslant C(|x - x_0|^2 + |y - y_0|^2)^{\alpha/2}$, where $0 \leqslant \alpha \leqslant 1$ is a Lipschitz exponent. If there exists $C > 0$ such that the previous condition is satisfied for any $(x, y) \in \Omega$, then f is uniformly Lipschitz α over the bounded domain Ω.

This property can be used as a predictor of visual smoothness or sharp intensity variation in the signal. The wavelet coefficients that correspond to smooth areas decrease rapidly as the resolution varies from coarse to fine, whereas a slow coefficient decay generally indicates the presence of either edges or detailed, textured areas. Formally, Lipschitz regularity is related to the asymptotic decay of $|W_s^k f(x, y)|$ $k = 1, \ldots, K$. This decay is controlled by the generalized WT modulus $\mathcal{W}_s f(x, y) = (\Sigma_{k=1}^{K} |W_s^k f(x, y)|^p)^{1/p}$, where usually[12] $p = 2$. Precisely, f is uniformly Lipschitz α over Ω if and only if there exists $A > 0$ such that, for all $(x, y) \in \Omega$ and all scales s, $|\mathcal{W}_s f(x, y)| \leqslant A s^{(\alpha+1)}$. If a partition function is defined[12]

$$Z(r, s) = \sum_{x,\, y} \mathcal{W}_s f(x, y)^r, \tag{23}$$

then the asymptotic decay is measured by the scaling index $\lambda(r)$,

$$\lambda(r) = \lim_{s \to 0} \frac{\log Z(r, s)}{\log s}. \tag{24}$$

Equation (24) implies that $Z(r, s) \sim s^{\lambda(r)}$. The algorithm to recover Lipschitz regularity at each point of the image computes $Z(r, s)$ over the modulus maxima at each scale and then reconstructs $\lambda(r)$ by means of linear regression.[12]

The method proposed in this paper is clearly related to the estimation of Lipschitz regularity. It is actually easy to note that the logarithm of the partition function $Z(r, s)$ as defined in Eq. (24) is equivalent, for $s = 2^l$, to Renyi's entropy defined in Eq. (16). In fact, considering Eq. (6), the coefficients w_l can be seen as a weighted modulus (for $p = 1$) of the $|w_l^k(n, m)|$ components. Our method, however, does not compute linear regression to characterize the behavior of the signal along the different scales; rather, it captures such a behavior by means of a majority vote strategy. Thus it can be seen as an approximate estimation of Lipschitz regularity. Clearly, here the goal is not to shape the Lipschitz regularity function precisely along scales but to take advantage of the different regularity of the target, as opposed to the background, for detecting one from the other.

Fig. 5. Example of detection obtained with the discussed optimal settings. From top to bottom: the original image, the truth image, and the final map.

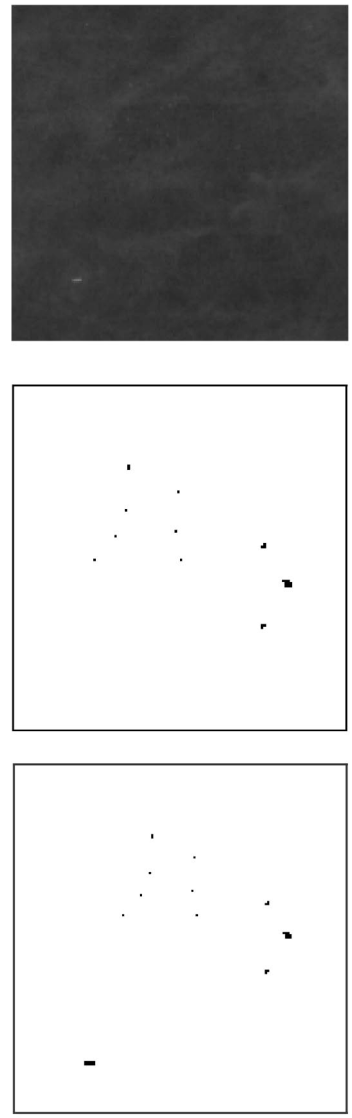

Fig. 6. Another example of detection.

This relation suggests an interesting interpretation. It is well known that the exponent $\lambda(r)$ is connected, by means of a Legendre transform, to the distribution or spectrum of singularities $D(\alpha)$.[12] This distribution, in turn, is defined as the fractal dimension of the set of all points characterized by a pointwise Lipschitz regularity α. If f is a multifractal signal (which indeed is a plausible assumption for the kind of images we are dealing with[22]), the spectrum $D(\alpha)$ measures the global repartition of singularities having different Lipschitz regularity. Basically, it gives the proportion of Lipschitz α singularities that appear at any scale s. On the other hand, it is well known that Renyi's information is a suitable tool for characterizing fractal dimensions,[20] each dimension being related to the order r of $I_{\Omega^p}(r)$. For example, we have derived Eq. (18) in the limit $r \rightarrow 2$: In this specific case the fractal dimension, evaluated by Renyi's information, corresponds to a correlation dimension[20] and indicates how the number of wavelet coefficient pairs, whose distance is less than s, changes with s. Note that such kind of second-order statistics has been usefully exploited in computational vision for capturing image textural properties.[23]

The fact that the image and, more precisely, that the background signal f_β can be modeled as a fractal noise sheds light on another facet of the correlation analysis performed by the WT. As previously discussed, we expect such correlation to be maximum when the wavelet is shaped like the searched target. This is the principle of the matched filter, which, in the case of fractal noise, can be discussed as follows.

A general fractal noise[20] is characterized by a generalized power spectrum that has a power decay $P_{f_\beta}(\omega_x) \sim 1/|\omega_x|^n$, where $n = 2H + 1$ and H is the Hurst parameter. H is related to the Lipschitz regularity α; for instance, if f_β is the realization of a fractional Brownian motion,[12] then it is almost everywhere singular with a pointwise Lipschitz regularity $\alpha = H$. The optimal procedure to detect f_τ is to apply a prewhitening matched filter $\hat{r}(\omega_x) = \hat{f}^*_\tau(\omega_x)/P_{f_\beta}(\omega_x)$, where $\hat{r}$ is the FT of the filter impulse response function $r(x)$ and $\hat{f}^*_\tau$ is the complex conjugate of the FT of the target signal $f_\tau(x)$. By using the $1/|\omega_x|^n$ power spectrum, one obtains

$$\hat{r}(\omega) = |\omega_x|^n \hat{f}^*_\tau(\omega_x). \tag{25}$$

For $n = 0$, Eq. (25) gives the classical matched filter of impulse response $r_0(x) = f^*_\tau(x)$ (appropriate for white flat spectrum noise). For $n \neq 0$, $\hat{r}(\omega_x) = (-i)^n (i|\omega_x|)^n \hat{r}_0(\omega_x)$ and, by applying the spatial differentiation property of the FT, $r(x) = (-i)^n (\partial^n/\partial x^n) r_0(x)$. Thus the prewhitening matched filter may be implemented by convolving the signal f with the filter $(\partial^n/\partial x^n) f^*_\tau(x)$, which consists of a generalized derivative operator followed by a normal matched filter, namely, $(\partial^n/\partial x^n)(f^*_\tau * f)(x)$. Taking into account that f_τ is real and scale dependent, we can choose $f^*_\tau(x)$ as $\bar{f}_{\tau,s}(x) = (1/\sqrt{s}) f_\tau(-x/s)$.

It has been proved[12] that a wavelet ψ with a fast decay has n vanishing moments if and only if there exists a function g with a fast decay such that $\psi(x) = (-1)^n [\partial^n g(x)/\partial x^n]$. As a consequence, the WT can be obtained as the convolution[12]

$$W_s f(u) = s^n \frac{\partial^n}{\partial x^n} (f * \bar{g}_s)(u), \tag{26}$$

with $\bar{g}_s(x) = (1/\sqrt{s}) g(-x/s)$. Then if $\bar{g}_s(x) = \bar{f}_{\tau,s}(x)$, Eq. (26) becomes $W_s f(x) = 2^n (\partial^n/\partial x^n)(f * \bar{f}_{\tau,s})(x)$; that is, by commuting convolution and differentiation, $W_s f(x) = s^n)(\partial^n/\partial x^n)(\bar{f}_{\tau,s} * f)(x)$. This shows that the WT acts as a prewhitening matched filter for detecting the target at scale s. Summing up, the wavelet is proportional to the generalized derivative of f_τ, namely, the signal we want to detect. Hence, the optimal detector is an admissible wavelet even if the initial template f_τ is not. This result includes the result of Strickland and Han's work, which presented evidence that, when f_τ is a Gaussian spot and the noise is one of the Markov types with power spectrum $O(|\omega|^{-2})$, the prewhitening matched filter is the second derivative of a Gaussian or a B-spline wavelet that closely approximates the second derivative of a Gaussian.

In this perspective, the best performance of a biorthogonal B-spline basis that we have experimentally reported with respect to other bases was not an unexpected result. The suitability of B-spline wavelets for processing images of the kind we are dealing with also stems from a different property, namely, the asymptotic convergence of B-spline wavelets to Gabor functions (modulated Gaussian).[24] Gabor functions are optimally concentrated in both the spatial and the spatial-frequency domains, which endows them with a remarkable aptitude for the analysis of nonstationary signals and textures. In fact, B-spline wavelets are the natural counterpart of the classic B spline. An essential property of these functions is their compact support, which makes them much better localized than their orthogonal equivalents, the Battle–Le Marie polynomial spline wavelets.

Eventually, it is of some value to make some connections of the topics discussed with some recent works in the scale-space literature. A first work by Sporring and Weickert[13] is concerned with entropy-based description of textures. To this end, a parametric image $f(t)$, where t is a parameter denoting diffusion time or scale, is transformed from fine to coarse scales by means of a first-order parabolic differential equation (diffusion equation). Namely, $(\partial f/\partial t) = \nabla^2 f$, ∇^2 denoting the Laplacian operator. Global information embedded in the image is traced through the change of Renyi's information by logarithmic scale,

$$c_r[f(t)] = \frac{\partial I_{\Omega f}(r)}{\partial \log t}, \tag{27}$$

where, clearly $I_{\Omega f}(r) = 1/(1 - r) \log \Sigma_{i=1}^N f_i^r(t)$. It is possible to show that such method shares some common points with our approach.

Consider again Eq. (26). Such a convolution defines a scale-space, and it is well known[12] that when the function g is the normalized Gaussian standard deviation $g(x, t) = 1/4\pi t \exp[-|x|^2/(4t)]$, the convolution (by setting $t = s$) provides a solution to the diffusion equation, g being the Green's function of the diffusion equation.[25] In other terms, the WT is proportional to a heat diffusion with initial condition $(\partial^n f)/(\partial x^n)$. In this case, the evaluation performed through Eq. (27) is clearly related to that provided by Eq. (24). However, the use of Renyi's entropy presented in this paper is specifically oriented to the detection problem, while Sporring and Weickert[13] are mainly concerned with a general discussion of information measures and a characterization of textural images by means of such measures.

A recent work by Ferraro *et al.*[14] allows us to compare the method proposed here with cross-entropy techniques in a multiscale framework. Cross-entropy modeling of the target detection task has been previously proposed and demonstrated in very simple situations[26] In contrast, in this work, owing to the employment of a thermodynamical model for scale-space generation, features are derived corresponding to entropy-rich regions of complex images. Relevant information contained in an image f (normalized as $\Sigma_{x,y \in D} f(x, y) = 1$) can be measured in terms of the distance between f and the corresponding fixed point f_∞, under the transformation T that takes the original image $f(\cdot, 0)$ to an image at a scale t, T_t:$f(\cdot, 0) \rightarrow f(\cdot, t)$. Such a distance measure is provided by the cross-entropy or Kullback–Leibler distance defined as[14]

$$H(f|f^*) = \sum_{x,y} f(x, y, t) \ln \frac{f(x, y, t)}{f_\infty(x, y)}, \tag{28}$$

where $f_\infty(\cdot)$ denotes the fixed point distribution of the transformation. On this basis, a dynamic measure of information is set up as the cross-entropy evolution $(\partial/\partial t) H(f|f_\infty)$, which can be measured during the transformation T. It has been shown that when an irreversible transformation (e.g., isotropic diffusion) is applied to an image, at any scale t, the minimization of $(\partial/\partial t) H(f|f_\infty)$ can measure the information loss of the process. Further, information can be gauged locally, at each point (x, y) of the image, through the density of entropy production $\sigma(x, y, t) = (\nabla f/f)^2$.

It can be noted that, by assuming a uniform distribution as that achieved in the case of isotropic diffusion, i.e., $f_\infty(x, y) = 1/P$, P being the dimension of the image domain, Eq. (28) can be reduced to

$$H(f|f^*) = \log P - \left[-\sum_{x,y} f(x, y, t) \log f(x, y, t) \right]. \tag{29}$$

The second term of Eq. (28) is Shannon's information, which can be derived as the Renyi's information of order 1, $I_{\Omega^p}(1)$. The expression basically asserts that, in the situation in which one of the distributions is uniform (say, f_∞), minimization of the cross-entropy between the two distributions is achieved by maximizing the entropy of the other distribution. Note that Eq. (18) together with Eq. (22) can be conceived as an information maximization-based technique.

Eventually, it is worth noting that the density of entropy production $\sigma(x, y, t)$ captures from a statistical physics standpoint the concept of signal regularity.[14] In fact, it is possible to measure the overall variation of $(\partial/\partial t) H(f|f_\infty)$ in terms of the σ activities along scales, the activity at (x, y) being defined as $\int \sigma(x, y, t) \mathrm{d}t$. On this basis, different activities allow us to discriminate between different types of image features (e.g., edges, textures, and uniform regions). These are encoded as regions of different information content within different bands of spatial scale. In the method proposed in this paper, majority voting provides an approximate, discrete evaluation of the coefficient activity along the scales of the WT. This is sufficient for our purposes, since our goal is not a classic segmentation of the image as proposed in the entropy production approach[14] but the detection of targets with respect to the background.

6. CONCLUSION

A method for the detection of small targets embedded within an inhomogeneous background was presented. The method relies upon thresholding the wavelet-transformed image to produce scale-dependent candidate detections, successively combined by means of majority voting.

The work reported presents various innovations. First, the method is based on a WT that is accomplished by means of a weighted, undecimated discrete scheme; a suitable tuning of the subband weights [Eq. (6)] endows such representation with a remarkable skill at capturing image characteristics. Also, the reconstruction problem is avoided, since the detection is performed in the transformed domain. Second, in contrast to other methods, our approach automatically determines the threshold, which we have demonstrated can be derived by maximizing the generalized Renyi's information [Eq. (18)]. Further, we have shown that an estimator close to the ideal wavelet coefficient selection can be achieved. This way, the correlation of information along scale can be reduced, for detection purposes, to a simple rule, namely, a majority vote rule.

Eventually, a physical interpretation of the method reveals interesting connections with results provided in different fields such as multifractal analysis, generalized information measures in scale-space, and cross-entropy analysis in fine-to-coarse transformations. This sheds new light on thresholding techniques that are usually conceived and discussed in a strict signal processing framework and, also on the more general links between information-theoretic methods performing in wavelet space and scale-space.

The theory was substantiated with experiments on real mammographic images, obtaining encouraging results. Experiments reported in Section 4 provide practical guidelines for researchers wishing to utilize the method for different kinds of images. In its essence, the approach does not assume any contextual information about the presence or absence of targets in the image; neither is an exact model of the background clutter and noise needed. Thus images with varying characteristics can be successfully segmented, avoiding the need for either fixed thresholds or specific *a priori* conditions. On the other hand, for effective use, the method requires some attention to be paid to the selection of the following parameters: the type of wavelet basis, the number of scales, and the γ weighting of the information carried in the subbands. All these parameters are application dependent.

The choice of wavelets and the range of scales are two well-known and crucial problems in the wavelet literature.[10,12] If one wants to avoid the self-construction of a wavelet basis, then attention should at least be paid to the number of vanishing points of the chosen wavelet, which is related to the order of differentiability and has a straightforward interpretation in terms of detectable structures. Once the wavelet basis has been selected, the range of scales can be derived by taking into account the size of the structure to be detected against the size of background structures. In the case studied here, it is easy to predict that most of the fine background structures will be matched at the finest scale, whereas they are discarded at subsequent, coarser scales $(L = 2, 3, 4)$. Nevertheless, it has to be remarked that the voting procedure has given evidence of appreciable flexibility with respect to this specific problem.

As regards the presented method, some particular attention should be devoted to the γ weighting of subband information. For this specific issue, ROC analysis, which is based on statistical decision theory, provides a reliable tool. We have employed it in the particular form of FROC curves, which constitutes a standard procedure in mammographic analysis. It is worth noting, however, that in general ROC curves provide objective estimates of the probabilities of decision outcomes (in terms of true-positive and false-positive decisions), and thus they be exploited for any and all of the criteria the system may have to tune.

ACKNOWLEDGMENTS

The authors are grateful to the referees, for their enlightening and valuable comments have greatly improved the quality and clarity of an earlier version of this paper.

The authors can be reached by e-mail at boccig @unisa.it and picus@unina.it.

REFERENCES

1. T. K. Bhattachary and S. Haykm, "ANN based adaptive radar detection scheme for small target in sea clutter," Electron. Lett. **28**, 1528–1536 (1992).
2. T. Soni, J. Zeidler, and W. Ku, "Performance evaluation of 2-D adaptive prediction filters for detection of small objects in image data," IEEE Trans. Image Process. **2**, 327–340 (1993).
3. P. French, J. Zeidler, and W. Ku, "Enhanced detectability of small objects in correlated clutter using an improved 2-D adaptive lattice algorithm," IEEE Trans. Image Process. **6**, 383–397 (1997).
4. R. N. Strickland and H. I. Han, "Wavelet transform for detecting microcalcifications in mammograms," IEEE Trans. Med. Imaging **15**, 218–229 (1996).
5. H. Chan, L. Niklason, D. Ikeda, K. Lam, and D. Adler, "Digitization requirements in mammography: effects on computer aided detection of microcalcifications," Med. Phys. **21**, 1203–1211 (1994).
6. I. Brodie and R. Gutcheck, "Radiographic information theory and application to mammography," Med. Phys. **9**, 812–820 (1982).
7. K. White, T. Hutson, and T. Hutchinson, "Modeling human eye behavior during mammographic scanning: preliminary results," IEEE Trans. Syst. Man Cybern. **27**, 494–505 (1997).
8. N. Karssemeijer, "Adaptive noise equalization and recognition of microcalcifications clusters in mammograms," Int. J. Pattern Recognition Artif. Intell. **7**, 1357–1377 (1993).
9. T. Cover and J. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
10. I. Daubechies, *Ten Lectures on Wavelets* (Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1992).
11. D. Donoho and I. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," Biometrika **81**, 425–455 (1994).
12. S. Mallat, *A Wavelet Tour of Signal Processing* (Academic, San Diego, Calif., 1998).
13. J. Sporring and J. Weickert, "Information measures in scale-spaces," IEEE Trans. Inf. Theory **45**, 1051–1058 (1999).
14. M. Ferraro, G. Boccignone, and T. Caelli, "On the representation of image structures via scale-space entropy condition," IEEE Trans. Pattern Anal. Mach. Intell. **21**, 1199–1203 (1999).
15. T. Kailath and H. V. Poor, "Detection of stochastic processes," IEEE Trans. Inf. Theory **44**, 2230–2259 (1998).
16. D. Gabor, "Theory of communication," J. IEE **93**, 429–457 (1946).
17. M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," IEEE Trans. Image Process. **1**, 205–220 (1992).
18. S. Mallat, "A theory of multiresolution signal decomposition: the wavelet representation," IEEE Trans. Pattern Anal. Mach. Intell. **11**, 674–693 (1989).
19. M. Shensa, "The discrete wavelet transform: wedding the à trous and Mallat algorithm," IEEE Trans. Signal Process. **40**, 2464–2482 (1992).
20. T. Takayasu, *Fractals in Physical Sciences* (Manchester U. Press, Manchester, UK, 1990).
21. J. Kittler, M. Hatef, R. B. W. Duin, and J. Matas, "On combining classifiers," IEEE Trans. Pattern Anal. Mach. Intell. **20**, 226–239 (1998).
22. L. Huai, K. Liu, and S. Lo, "Fractal modeling and segmentation for the enhancement of microcalcifications in digital mammograms," IEEE Trans. Med. Imaging **16**, 785–797 (1997).
23. B. Wandell, *Foundations of Vision* (Sinauer, Sunderland, Mass., 1995).
24. M. Unser, A. Aldroubi, and M. Eden, "On the asymptotic convergence of B-spline wavelets to Gabor functions," IEEE Trans. Inf. Theory **38**, 864–872 (1992).
25. P. Morse and H. Feshbach, *Methods of Theoretical Physics* (McGraw-Hill, Toronto, 1953).
26. D. Sheffer and D. Ingman, "The informational difference concept in analyzing target recognition issues," J. Opt. Soc. Am. A **14**, 1431–1438 (1997).