Wavelets for a Vision

STÉPHANE MALLAT

Invited Paper

Early on, computer vision researchers have realized that multiscale transforms are important to analyze the information content of images. The wavelet theory gives a stable mathematical foundation to understand the properties of such multiscale algorithms. This tutorial describes major applications to multiresolution search, multiscale edge detection, and texture discrimination.

I. Introduction

Multiscale processing is hardly avoidable to develop efficient image recognition algorithms. Before wavelets were called "wavelets," researchers such as Burt and Adelson [7], Koenderink [18], Marr [24], Witkin [36], and Rosenfeld [30] had established the necessity to extract multiscale image information. Some of these ideas have later been formalized and refined by the wavelet theory. In parallel, psychophysics, and physiological experiments [11] have shown that multiscale transforms seem to appear in the visual cortex of mammals. This was an important motivation to further study the application of such transforms to image analysis. To explain the impact of wavelets for low-level vision, we concentrate on three major applications: multiresolution processing, multiscale edge detection, and texture discrimination.

Multiresolution algorithms modify the image resolution to process as little data as possible, for any particular visual task. Coarse to fine searches process first a low resolution image and zoom selectively into finer scale information, if necessary. Applications to stereo vision and optical flow measurements are described.

Local image contrasts are often more informative than light intensity values. A wavelet transform measures gray level image variations at different scales. Contours of image structures correspond to sharp contrasts and can be detected from the local maxima of a wavelet transform. Their importance is illustrated by our ability to recognize complex scenes from a drawing that outlines edges. The wavelet theory relates the behavior of multiscale edges to local image properties. It also opens the door to reconstruction algorithms which recover images from multiscale edges.

Among low-level vision problems, texture discrimination is certainly one of the most difficult. Despite the fact that textures are quickly preattentively discriminated by a human observer [21], there is still no appropriate model for textures. The perception of textures as opposed to edges depends upon local but not pointwise properties. However, there is no predefined neighborhood size over which textures can be analyzed. This has motivated the use of wavelet transforms that measure the image properties over domains of varying sizes. Local frequency measurements derived from a directional wavelet transform appear to be important for texture discrimination [5], [15], [29]. Yet no comprehensive theory guides texture segmentations from wavelet coefficients.

When studying the application of wavelets to computer vision, the major difficulties arise at the interface between low-level algorithms and higher level visual models. Multiresolution search strategies must depend upon prior knowledge on the world. Similarly, edges detection can not be restricted to a pointwise processing as it shown by our perception of illusory contours [19]. Texture discrimination also requires the elaboration of prior models which guide the grouping procedures for image segmentations. We discuss these issues in more details.

II. Multiresolution Processing

B. Fovea and Multiresolution Pyramids

An image of 512 × 512 pixels often includes too much information for real time vision processing. Multiresolution algorithms process less image data by selecting the relevant details that are necessary to perform a particular recognition task. The human visual system uses a similar strategy. The distribution of photoreceptors on the retina is not uniform. The visual acuity is the greatest at the center of the retina where the density of receptors is maximum. When moving away from the center, the resolution decreases proportionally to the distance from the retina center

Manuscript received June 1, 1995; revised December 15, 1995. This work was supported by the AFOSR grant F49620-93-1-0102 and ONR grant N00014-91-J-1967.

The author is with the Courant Institute of Mathematical Sciences, New York University, New York, NY 10003 USA, and also with Ecole Polytechnique, Department of Applied Mathematics, 91128 Palaiseau, France.

Publisher Item Identifier S 0018-9219(96)02576-5.

[33]. The high resolution visual center is called fovea. It is responsible for high acuity tasks such as reading or recognition. A retina with a uniform resolution equal to the highest fovea resolution would require about 10 000 times more photoreceptors. Such a uniform resolution retina would increase considerably the size of the optic nerve that transmits the retina information to the visual cortex and the size of the visual cortex that processes this data.

Active vision [1] strategies compensate the nonuniformity of visual resolution by moving the fovea with eye saccades. Regions of a scene with a high information content are scanned successively. These saccades are partly guided by the lower resolution information gathered at the periphery of the retina. This multiresolution sensor has the advantage to provide high resolution information at selected locations and a large field of view with relatively little data.

Multiresolution algorithms implement in software the search for important high resolution data. A uniform high resolution image is measured by a camera but a small part of this information is processed. The high resolution information is selectively considered depending upon lower resolution processing [6]. Such algorithms are efficiently implemented with multiresolution pyramids introduced by Burt and Adelson [7].

Let us normalize the image resolution to one. A multiresolution pyramid computes the image approximation at lower resolutions 2^j for $j < 0$. As explained in the background article [10], an approximation of $f(x,y)$ at a resolution 2^j is defined as an orthogonal projection on a space $\boldsymbol{V}_j$. It has been proved that such multiresolution spaces admit orthogonal basis of $\boldsymbol{V}_j$ of dilated separable scaling functions

$$\{\sqrt{2^j}\phi(2^jx-n)\sqrt{2^j}\phi(2^jy-m) \\ = \phi_{j,n}(x)\phi_{j,m}(y)\}_{(n,m)\in\boldsymbol{Z}^2}.$$

The approximation at a resolution 2^j is thus characterized by the inner products

$$f^j[n,m] = \langle f(x,y), \phi_{j,n}(x)\phi_{j,m}(y)\rangle.$$

We suppose that $f^0[n,m]$ is the discrete image at the resolution one measured by the camera. One can prove that image approximations $f^j[n,m]$ at smaller resolutions are computed with a succession of low-pass filterings and subsamplings [22]. Let $h[n]$ be the Conjugate Mirror Filter associated to the scaling function $\phi(t)$ and $h_2[n,m] = h[-n]h[-m]$. An image $f^j[n,m]$ at a resolution 2^j is obtained from a higher resolution image $f^{j+1}[n,m]$ with a low-pass filtering with $h_2[n,m]$ and a subsampling by two along the rows and columns

$$f^j[n,m] = f^{j+1} \star h_2[2n,2m].$$

If $f^0[n,m]$ has N^2 nonzero samples, with appropriate border treatments, $f^j[n,m]$ has $2^{2j}N^2$ nonzero pixels. Fig. 1 shows an example of multiresolution image pyramid over five octaves.

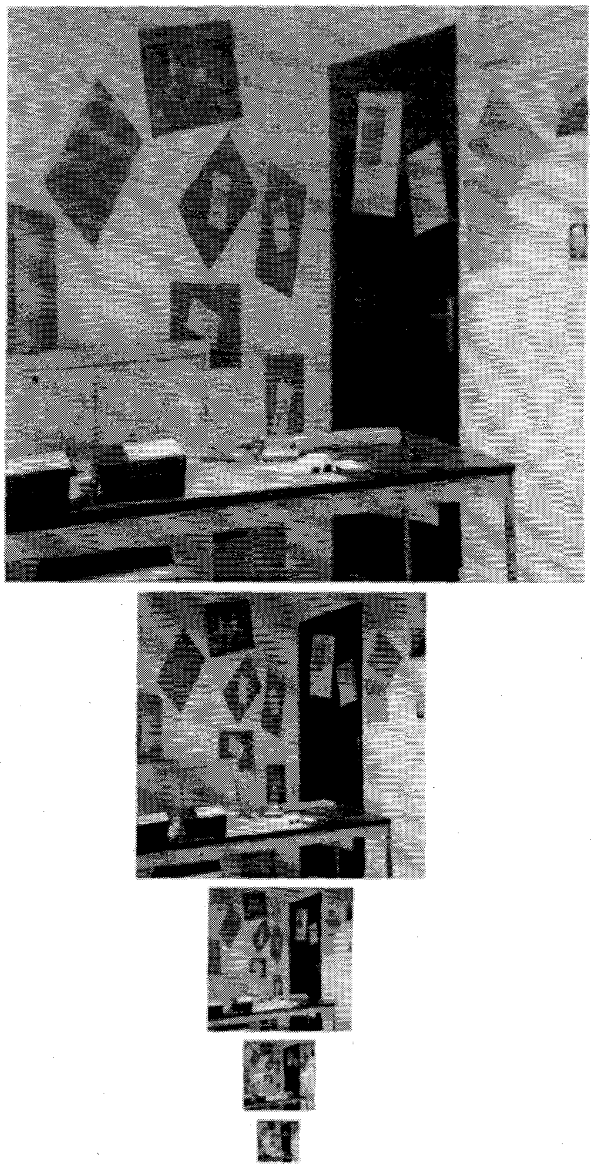


Fig. 1. Multiresolution image pyramid. From top to bottom the resolution decreases by two from one image to the next. These images are obtained with a cascade of low-pass filtering and subsampling.

B. Coarse to Fine Multiresolution Processing

Coarse to fine multiresolution search reduces the computational complexity by beginning at low-resolution and adaptively increasing the resolution to gather the necessary details. We describe applications to the estimation of optical flow from time image sequences and depth from stereo images.

The optical flow is computed from a sequence of images at time intervals Δ. Let $I_k[n,m]$ be the gray level image intensity at time $k\Delta$. A pixel (n_0,m_0) gives the light intensity reflected by a point P in the 3-D scene. If the image gray level is not constant in the neighborhood of (n_0,m_0), a change between the relative position of P and the camera creates an intensity displacement. The velocity of this gray level displacement in the image plane is called the optical flow. If there is no change of lightning, it can be related to the 3-D velocity of P [9]. There are several approaches to compute the optical flow, including the use of local differential operators [9]. A simple technique is to estimate the gray level displacement from frame to frame by finding a correspondence between the pixels of successive images $I_k[n,m]$ and $I_{k+1}[n,m]$.

A similar point matching problem appears in stereo vision [12]. A point P in a 3-D scene is projected at different locations in the image planes of two stereo cameras. From the difference between the position of this projection in the left and right images of a stereo pair, one can recover the 3-D coordinates of P in a referential related to the stereo cameras. The main difficulty of stereo vision is to match each pixel of the left image $I_l[n,m]$ to the pixel of the right image $I_r[n,m]$ which is the projection of the same point P in the scene.

To find the correspondence between pixels of two images $I_1[n,m]$ and $I_2[n,m]$, one can maximize a correlation measure. We search for a point (n_2,m_2) in $I_2[n,m]$ whose neighborhood of size K has a maximum correlation with the neighborhood of a pixel (n_1,m_1) in $I_1[n,m]$. The normalized correlation is defined by (see (1) at the bottom of the page).

This correlation is always smaller than one and is equal to one if and only if the neighborhood of (n_2,m_2) in $I_2[n,m]$ is proportional to the neighborhood of (n_1,m_1) in $I_1[n,m]$.

If $I_1[n,m]$ and $I_2[n,m]$ have N^2 pixels, this exhaustive correlation search requires K^2N^4 multiplications, which is huge. Another difficulty is to find an appropriate template size K. If K is too small, the neighborhood

$$\{I_1[n_1-n, m_1-m]\}_{-(K/2)\le n,m\le(K/2)}$$

of (n_1,m_1) might not contain enough information to disambiguate several potential matches in the image $I_2[n,m]$. If K is too large there might not be any appropriate match in $I_2[n,m]$. This is the case for optical flow measurements if the neighborhood $\{I_1[n_1-n,m_1-m]\}_{-(K/2)\le n,m\le(K/2)}$ includes smaller components having different displacements [3]. The overall region is not globally translated and thus does not match well any other domain of the next image. In stereo vision, the same problem appears if the neighborhood of size K includes objects whose distance to the camera are very different [12]. The perspective projection then induces important distortions between the projections on the left and right camera planes. The difficulty to choose an appropriate neighborhood size and the large computational complexity motivates the use of multiresolution correlations.

A multiresolution matching algorithm correlates first low resolution approximations of $I_1[n,m]$ and $I_2[n,m]$ and refines the match at high resolution guided by the lower resolution estimates [31]. Let us compute the multiresolution pyramids $\{I_1^j[n,m]\}_{J\le j\le -1}$ and $\{I_2^j[n,m]\}_{J\le j\le -1}$ of $I_1[n,m]$ and $I_2[n,m]$, with a maximum depth $-J\le \log_2 N$. We first correlate the points of the lower resolution images $I_1^J[n,m]$ and $I_2^J[n,m]$ over neighborhoods of size K, which is typically equal to three or five. For any point (n_1^J,m_1^J) of $I_1^J[n,m]$ we find (n_2^J,m_2^J) in $I_2^J[n,m]$ whose neighborhood maximizes the normalized correlation (1). Since I_1^J and I_2^J have only $2^{2J}N^2$ pixels $(J<0)$, this correlation is performed with much fewer operations than on the original image. This low resolution matching is used to constrain to a limited area the correlation search at the next resolution. The region around (n_1^J,m_1^J) in I_1^J corresponds in the image $I_1^{J+1}[n,m]$ to a region around one of the points

$$\begin{aligned}(n_1^{J+1},m_1^{J+1})\in\{&(2n_1^J,2m_1^J),(2n_1^J+1,2m_1^J),\\ &(2n_1^J,2m_1^J+1),(2n_1^J+1,2m_1^J+1)\}\end{aligned}$$

(see Fig. 2). Similarly, (n_2^J,m_2^J) in $I_2^j[n,m]$ corresponds to a location close to $(2n_2^J,2m_2^J)$ in $I_2^{J+1}[n,m]$. At the resolution 2^{J+1}, we correlate a square neighborhood of (n_1^{J+1},m_1^{J+1}) of size K in $I_1^{J+1}[n,m]$ with neighborhoods in $I_2^{J+1}[n,m]$ centered at locations close to $(2n_2^J,2m_2^J)$. The center location (n_2^{J+1},m_2^{J+1}) which maximizes the normalized correlation (1) is a higher resolution match of (n_1^{J+1},m_1^{J+1}). The resolution of the matching is progressively refined with the same procedure from one resolution to the next, until the finest resolution $2^j=1$.

A single pixel at a resolution 2^j $(j<0)$ covers a block of 2^{-j} pixels in the original image at the resolution 1 (see Fig. 2). A correlation with a template of size K on $I_2^j[n,m]$ is thus equivalent to a correlation over a neighborhood of size $2^{-j}K$ in the original image $I_2[n,m]$. By letting constant the size K of the correlation templates at all resolutions, the algorithm performs correlations over neighborhoods whose size vary proportionally to 2^{-j} relatively to the original image. The coarse information is correlated over large neighborhoods whereas the fine information is correlated over small neighborhoods. Varying this size avoids choosing between small templates that might not disambiguate a match and large templates that may produce wrong match if the images $I_1[n,m]$ and $I_2[n,m]$ are

$$\frac{\displaystyle\sum_{n,m=-(K/2)}^{K/2} I_1[n-n_1,m-m_1]I_2[n-n_2,m-m_2]}{\left(\displaystyle\sum_{n,m=-(K/2)}^{K/2}|I_1[n-n_1,m-m_1]|^2\right)^{1/2}\left(\displaystyle\sum_{n,m=-(K/2)}^{K/2}|I_2[n-n_2,m-m_2]|^2\right)^{1/2}} \quad (1)$$

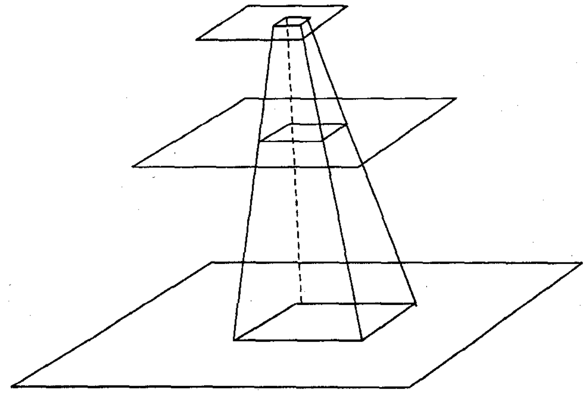


Fig. 2. A square neighborhood of width K at low resolution 2^j corresponds to square neighborhoods of size $2K$ and $4K$ at higher resolutions 2^{j+1} and 2^{j+2}.

locally warped. When estimating an optical flow at a coarse resolution 2^j, we mentioned that there might be regions of size 2^jK in $I_1[n,m]$ which include components with different velocities. These structures are translated at different positions of $I_2[n,m]$, which modifies the properties of gray level neighborhoods. In this case, a correlation of the coarse resolution images $I_1^j[n,m]$ and $I_2^j[n,m]$ produces a wrong estimate of the flow. This error can be detected at a higher resolution 2^l where the smaller correlation length $2^{-l}K$ can resolve the regions having different motions. It is thus necessary to use verification strategies that detect misleading coarse resolution information from the finer resolution data [3].

The main difficulty when implementing a multiresolution algorithm is to find efficient strategies to select the high resolution information. The potential errors induced by low resolution processing requires the use of verification procedures which incorporate high resolution verifications to guide the search. Such algorithms have already been developed for optical flow [3] and stereo vision [12]. For pattern recognition, the problem is more difficult because the search must also be guided with prior information about potential patterns in the scene. For example, to recognize a person from a photograph, the visual saccades of a human observer concentrate mostly the attention of the fovea on the eyes of the face in the photograph [1]. The eyes provide important cues for recognition and clearly the saccades must have incorporated high level information about faces to derive this strategy. The integration of such high level information to guide automatically the multiresolution data search is still an open problem.

C. Why Wavelet Bases are Not Used

The background article [10] shows that wavelet bases extract the necessary information to increase the resolution of an image approximation. One would thus expect that these bases can play an important role in multiresolution visual processing. The sad reality is that wavelet bases have not yet found any application for visual pattern recognition, because of their lack of translation invariance. To simplify the explanations, we describe the problem for 1-D signals. In a basis, wavelet coefficients at a scale 2^j are inner products with wavelets dilated by 2^j and translated by 2^jn

$$\langle f, \psi_{j,n}\rangle = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{2^j}}\psi\left(\frac{t-2^jn}{2^j}\right)dt = Tf(2^j, 2^jn)$$

with a continuous wavelet transform defined by

$$Tf(a,b) = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right)dt.$$

The wavelet basis coefficients are thus obtained by sampling uniformly the continuous wavelet transform at dyadic scales $\{2^j\}_{j\in\mathbf{Z}}$, and at time locations 2^jn proportional to the scale. If $f(t)$ is translated by τ the continuous wavelet transform of $f_\tau(t) = f(t-\tau)$ is translated by the same amount

$$Tf_\tau(2^j, b) = Tf(2^j, b-\tau).$$

However, the sampled coefficients $\{Tf_\tau(2^j, 2^jn) = Tf(2^j, 2^jn-\tau)\}_{n\in\mathbf{Z}}$ are not equal to a translation of the values $\{Tf(2^j, 2^jn)\}_{n\in\mathbf{Z}}$, when τ is not proportional to 2^j. As a result, the wavelet coefficients of a translated function $f_\tau(t)$ may be very different from the wavelet coefficients of $f(t)$. It is difficult to characterize a pattern from the wavelet coefficients in a basis since these wavelet descriptors depend upon the pattern location.

III. Multiscale Edge Detection

The evocative power of drawings clearly shows that edges are among the most important features for pattern recognition. But what is an edge? When looking at a brick wall, we may decide that the edges are the contours of the wall whereas the bricks define a texture. We may also include the contours of each brick among the set of edges and consider the irregular surface of each brick as a texture. The discrimination of edges variations versus textures depends upon the scale of analysis. This has motivated computer vision researchers to detect sharp image variations at different scales [24], [30], [36].

The next section describes a multiscale Canny [8] edge detector that is most often used in vision algorithms. This edge detector is equivalent to the detection of wavelet transform local maxima. The wavelet theory allows one to understand how to combine multiscale edge information to characterize different types of edges. It also provides the mathematical grounds to implement an algorithm that reconstructs images from edges.

A. Wavelet Maxima

Canny's algorithm [8] detects sharp variation points of an image $f(x,y)$ from the modulus of the gradient vector

$$\vec{\nabla} f(x,y) = \begin{pmatrix} \dfrac{\partial f(x,y)}{\partial x} \\ \dfrac{\partial f(x,y)}{\partial y} \end{pmatrix}.$$

The partial derivative of $f(x,y)$ in a direction $\vec{n}$ of the (x,y) plane is equal to the inner product

$$\frac{\partial f(x,y)}{\partial n} = \vec{\nabla} f(x,y) \cdot \vec{n}.$$

The absolute value of this partial derivative is maximum if $\vec{n}$ is parallel to $\vec{\nabla} f$. This proves that the gradient vector points locally in the direction of maximum change of the surface. A point (x_0, y_0) is defined to be an edge point if the modulus of $\vec{\nabla} f(x,y)$ is locally maximum at (x_0, y_0), when (x,y) varies in a 1-D neighborhood of (x_0, y_0) that is collinear to the direction of $\vec{\nabla} f(x_0, y_0)$. These edge points are locations where the surface has locally a maximum rate of change. They are inflection points of $f(x,y)$.

A multiscale version of this edge detector is implemented by smoothing the surface with a convolution kernel $\theta(x,y)$ that is dilated at dyadic scales $\{2^j\}_{j \in \boldsymbol{Z}}$. Such an edge detector can be computed with two wavelets that are the partial derivatives of $\theta(x,y)$

$$\psi^1(x,y) = \frac{\partial \theta(x,y)}{\partial x}, \quad \psi^2(x,y) = \frac{\partial \theta(x,y)}{\partial y}. \tag{2}$$

Let us denote

$$\psi^k_{2^j}(x,y) = \frac{1}{2^j} \psi^k\left(\frac{x}{2^j}, \frac{y}{2^j}\right) \quad \text{for} \quad 1 \le k \le 2.$$

The wavelet transform of $f(x,y)$ at a scale 2^j has two components which can be written as convolutions with $\tilde{\psi}^k_{2^j}(x,y) = \psi^k_{2^j}(-x,-y)$

$$\begin{aligned} T^k f(2^j, u, v) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y) \psi^k_{2^j}(x-u, y-v) \\ &\quad \cdot dx\, dy \\ &= f \star \tilde{\psi}^k_{2^j}(u,v). \end{aligned} \tag{3}$$

If we suppose that there exists $A > 0$ and $B \ge A$ such that the Fourier transforms $\hat{\psi}^k(\omega_x, \omega_y)$ of these wavelets satisfy

$$\forall (\omega_x, \omega_y) \in \boldsymbol{R}^2, \quad A \le \sum_{k=1}^{2} \sum_{j=-\infty}^{+\infty} |\hat{\psi}^k(2^j \omega_x, 2^j \omega_y)|^2 \le B \tag{4}$$

one can then prove that the 2-D dyadic wavelet transform (3) defines a complete and stable signal representation.

Let us denote

$$\tilde{\theta}_{2^j}(x,y) = \frac{1}{2^j} \theta\left(\frac{-x}{2^j}, \frac{-y}{2^j}\right).$$

The two wavelets can be rewritten

$$\begin{aligned} \tilde{\psi}^1_{2^j}(x,y) &= -2^j \frac{\partial \tilde{\theta}_{2^j}(x,y)}{\partial x} \\ \text{and} \quad \tilde{\psi}^2_{2^j}(x,y) &= -2^j \frac{\partial \tilde{\theta}_{2^j}(x,y)}{\partial y}. \end{aligned} \tag{5}$$

Inserting (5) in (3) and putting the partial derivative outside the convolution products proves that

$$\begin{aligned} \begin{pmatrix} T^1 f(2^j, u, v) \\ T^2 f(2^j, u, v) \end{pmatrix} &= -2^j \begin{pmatrix} \frac{\partial}{\partial u}(f \star \tilde{\theta}_{2^j})(u,v) \\ \frac{\partial}{\partial v}(f \star \tilde{\theta}_{2^j})(u,v) \end{pmatrix} \\ &= -2^j \vec{\nabla}(f \star \tilde{\theta}_{2^j})(u,v). \end{aligned} \tag{6}$$

The two components of the wavelet transform are proportional to the coordinates of the gradient vector of $f(x,y)$ smoothed by $\tilde{\theta}_{2^j}(x,y)$. The modulus of the gradient vector $\vec{\nabla}(f \star \tilde{\theta}_{2^j})(u,v)$ is thus proportional to the wavelet transform modulus

$$Mf(2^j, u, v) = \sqrt{|T^1 f(2^j, u, v)|^2 + |T^2 f(2^j, u, v)|^2} \tag{7}$$

and its angle is

$$Af(2^j, u, v) = \operatorname{argtan}\left(\frac{T^2 f(2^j, u, v)}{T^1 f(2^j, u, v)}\right).$$

Following Canny's approach [8], the edges at the scale 2^j are defined as points (u_0, v_0) where $Mf(2^j, u, v)$ is locally maximum in the 1-D neighborhood that is along the direction given by $Af(2^j, u, v)$. These points are also called wavelet transform modulus maxima. As opposed to wavelet basis coefficients, when the image is translated, the wavelet maxima are translated without being modified. This property is particularly important for the application of multiscale edges to pattern characterization.

The first column of Fig. 3 displays the wavelet transform modulus $Mf(2^j, u, v)$ over four octaves of the image in Fig. 1. Dark pixels correspond to high amplitude modulus points. The second column gives the angle $Af(2^j, u, v)$ which varies from zero (white) to 2π (black). When the modulus is close to zero, the angle measurement is unstable and is set to zero. This wavelet transform is computed with a compactly supported window $\theta(x,y)$ that is a separable product of cubic spline functions. If the original image has N^2 pixels, the wavelet transform is computed over $\log_2 N$ scales with $O(N^2 \log_2 N)$ operations, by using a filter bank algorithm [23]. The wavelet maxima are displayed in the third column. At fine scales, there are many edge points created by the image noise. Most of these maxima are removed by the smoothing at larger scales. However, this smoothing also changes the location of edges.

Edge points are distributed along curves in the image plane that often correspond to the boundary of important structures. To recover these edges curves, individual wavelet modulus maxima are chained. Since the gradient vector points in the direction of maximum change of the intensity surface, the angle $Af(2^j, u, v)$ is orthogonal to the tangent of the edge curve of $f \star \tilde{\theta}_{2^j}(x,y)$ that goes through (u,v). In discrete computations, we chain two wavelet modulus maxima that are neighbors if the vector that joins these two points is perpendicular to the angle direction $Af(2^j, u, v)$ at these points. The fourth column of Fig. 3 displays the edges chains that include more than 10 pixels, and along which the average modulus value $Mf(2^j, u, v)$ is larger than a specified threshold. Small edges created by noises are removed by this chain thresholding.

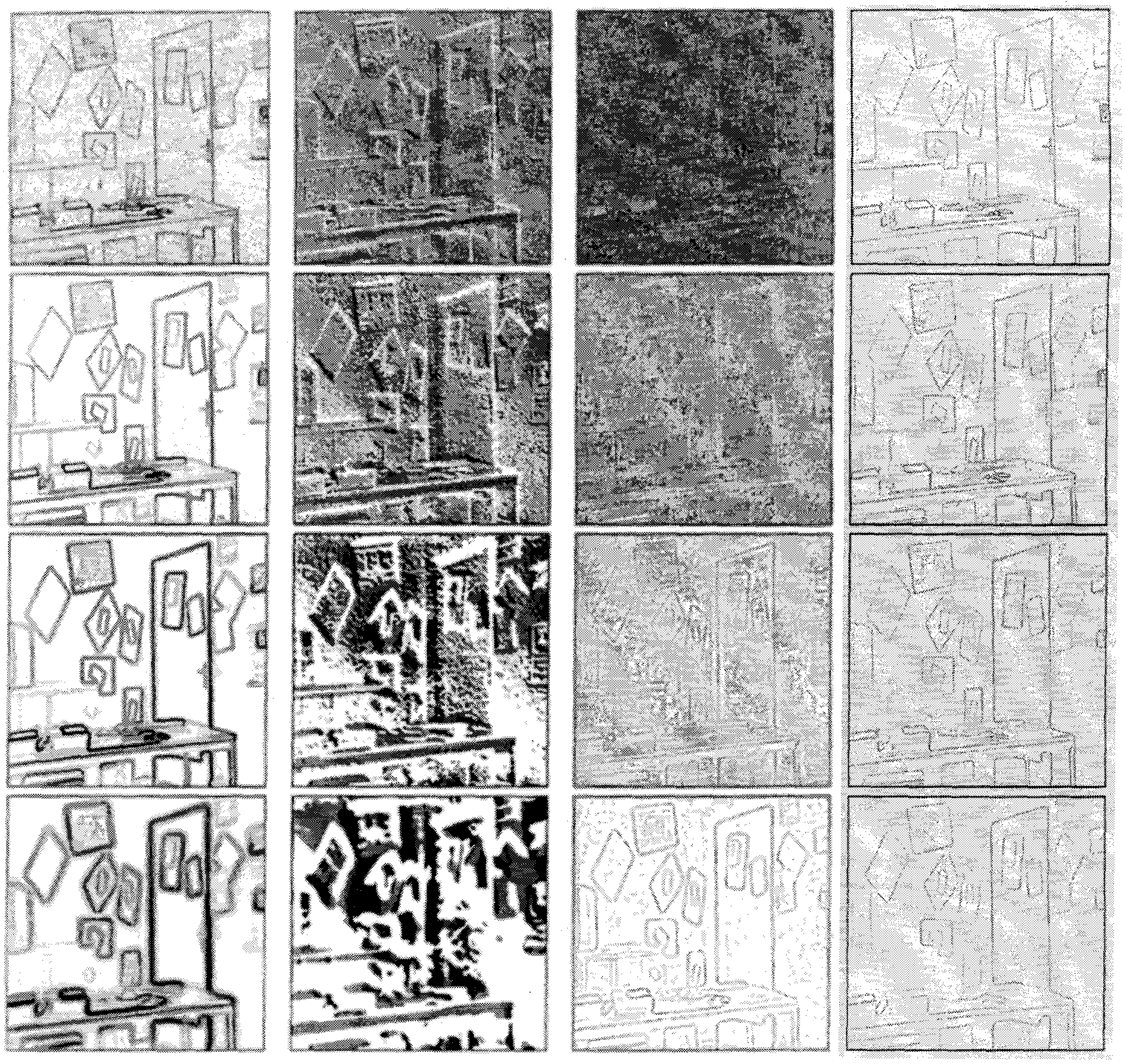


Fig. 3. The original image is at the top left of Fig. 4. The first and second columns display the wavelet modulus $Mf(2^j,u,v)$ and angles $Af(2^j,u,v)$ at scales 2^j for $1 \leq j \leq 4$. The pixels darkness are proportional to the amplitude of $Mf(2^j,u,v)$ and $Af(2^j,u,v)$ (which goes from zero to 2π). The black pixels of the images on the third column are the modulus maxima of $Mf(2^j,u,v)$ along the direction specified by $Af(2^j,u,v)$. These maxima are chained together. The fourth columns displays the longer edge chains with higher average modulus values.

B. Multiscale Edge Processing

Once edges are detected at several scales, we must understand how to integrate this multiscale information for pattern recognition. One might be tempted to look for a "best" scale where the edges are well discriminated from noises and textures. The wavelet theory shows that much finer properties are derived by analyzing edge behaviors across scales. The multiscale edge information is in fact rich enough to recover close image approximations.

The background article [10] explains that the decay of a wavelet transform depends upon the local regularity of the signal. This regularity is quantified by Lipschitz exponents. A function $f(x,y)$ is said to be Lipschitz α at (x_0,y_0), with $0 \leq \alpha \leq 1$, if for all points (x,y) in a 2-D neighborhood of (x_0,y_0)

$$|f(x,y) - f(x_0,y_0)| \leq K(|x-x_0|^2 + |y-y_0|^2)^{\alpha/2}. \quad (8)$$

The larger α, the more regular the function. At a discontinuity, the function is Lipschitz $\alpha = 0$. If $1 > \alpha > 0$ the image is continuous but not differentiable and α characterizes the type of singularity at that location. The Lipschitz regularity of a function $f(x,y)$ is related to the asymptotic decay of the two wavelet components $|T^1 f(2^j,u,v)|$ and $|T^2 f(2^j,u,v)|$ when the scale 2^j decreases. This decay is controlled by the modulus $Mf(2^j,u,v)$ and one can prove [25] that a necessary condition for $f(x,y)$ to be Lipschitz α at (x_0,y_0) is the existence of $C > 0$ such that

$$|Mf(2^j,u,v)| \leq C2^{j(\alpha+1)}. \quad (9)$$

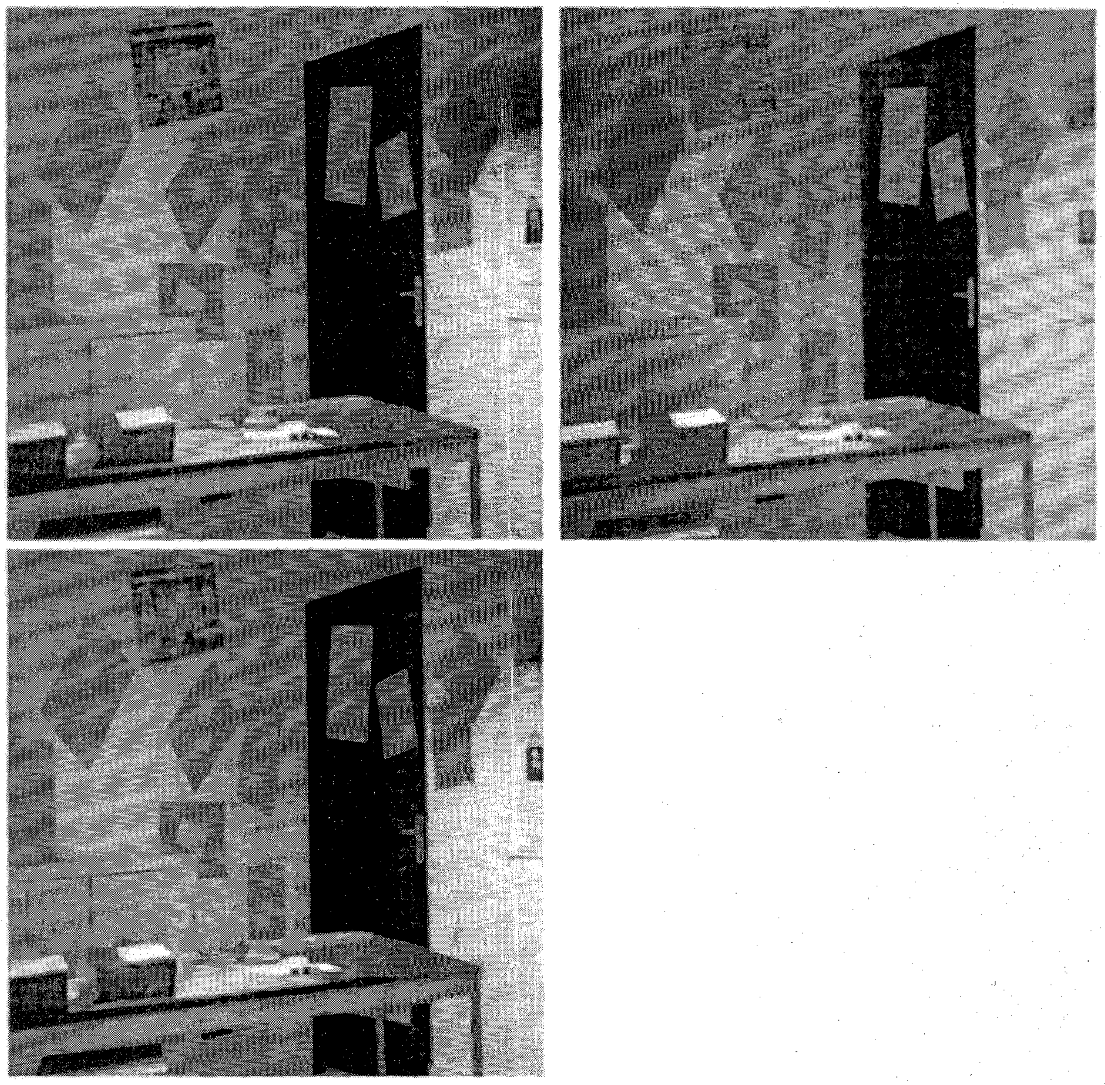


Fig. 4. The image at the top left is the original image. The image at the top right is reconstructed from the multiscale wavelet maxima shown in the third column of Fig. 3, plus the lower scale information. The image at the lower right is reconstructed from the thresholded edges shown in the fourth column of Fig. 3.

Suppose that the image has an isolated edge curve along which $f(x, y)$ has singularities which are Lipschitz α. By measuring the decay of the modulus maxima across scales, we derive from (9) an estimate of the Lipschitz regularity along the edge. At some edge locations, the image is not singular but has a smooth transition that is locally sharper. For example, the diffraction effect creates smooth edges at the borders of shadows. An analysis of the decay of wavelet maxima can also provide an estimate of the local smoothness of edges [23].

Multiscale edges give a rich description of the image information and one may wonder whether it is possible to reconstruct the whole image from these edges. This issue was raised by Marr [24] and studied by several researchers in computer vision [14], [37]. The wavelet theory allows one to express the nonlinear constraints derived from the knowledge of the modulus maxima locations (u_n, v_n) as well as the values of $Mf(2^j, u_n, v_n)$ and $Af(2^j, u_n, v_n)$ at these locations. An alternate projection algorithm recovers an image which belongs to the set of functions whose wavelet transform satisfies these maxima constraints [23]. The upper right image of Fig. 4 shows the reconstructed image from the multiscale edges displayed in Fig. 3. Since edges are computed up to the scale 2^4, the image low-frequencies at scales larger than 2^4 are used to complement the edge information in the reconstruction. When edges are computed up to the coarser scale $\log_2 N$, this complement of information is reduced to the average value of the image intensity. Extensive numerical experiments show that the reconstructed images are visually identical to the original ones, although Meyer [26] and Berman [4] proved that an image is not uniquely characterized by its multiscale edges. There are no visual distortions because reconstruction errors remain below the visual sensitivity threshold. The mathematical problem is still open and despite further studies [35], we do not understand why these reconstruction algorithms work so well and under what condition multiscale edges do provide a complete and stable signal representation. Since we can reconstruct visually perfect images from multiscale edges, one can develop image processing algorithms that manipulate the image information over the edge representation. Some edges and singularities can be suppressed from the image by suppressing wavelet maxima and reconstructing the corresponding image. The lower left image in Fig. 4 shows the image reconstructed from the thresholded multiscale edges displayed in the right column of Fig. 3. This reconstructed image is nearly identical to the original one because the small edges that have been suppressed mostly correspond to noise variations. Applications to noise removal and compact signal coding have been developed [23].

A multiscale wavelet edge detector defines edges as points where the image intensity has sharp variations. This definition is however too restrictive when edges are used to find the contours of objects. For image segmentations, edges must define closed curves that outline the boundaries of each region. Because of noise or light variations, generally there are holes in the contours obtained by a local edge detector. Filling these holes requires some prior knowledge on the expected behavior of edges in the image. The illusions of the Kanizsa triangles [17] clearly show such an edge "filling" is performed by the human visual system. The illusion gives us the impression that there exists an edge at locations where the image intensity is constant, in order to close the contour of a triangle region. Closing edge curves and understanding illusory contours requires computational models that are not as local as multiscale differential operators. Variational approaches that incorporate the expected regularity of the contours give promising strategies to close contours and understand the perception of illusory contours [19].

IV. Texture Discrimination

A texture segmentation divides the image into "homogeneous" regions where local texture properties are approximatively invariant. Despite many attempts, there is still no appropriate model for "homogeneous textures." Right now, a texture homogeneity is defined only with respect to our visual perception. A region is said to have a homogeneous texture if it is preattentively perceived as being homogeneous by a human observer.

Rather than constructing formal models, several algorithmic approaches have tried to isolate important texture parameters for recognition. These ideas are often inspired by the "texton" theory of Julesz [16] who searched for elementary patterns which are responsible for our discrimination abilities. The goal is to find a minimum number of measurements that can discriminate textures that are perceived to be "different." These measurements should also remain approximatively constant in a region where the texture is considered to be homogeneous.

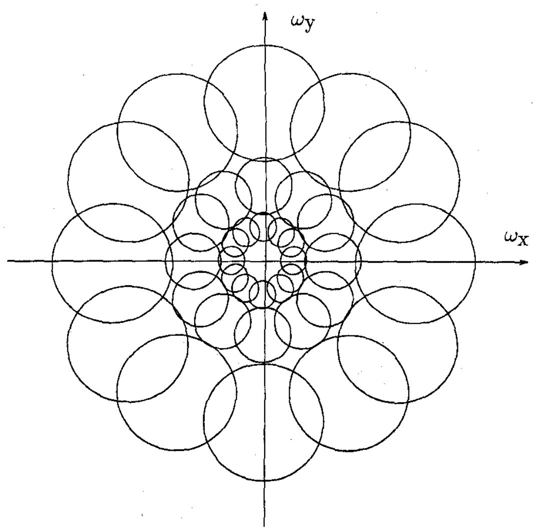
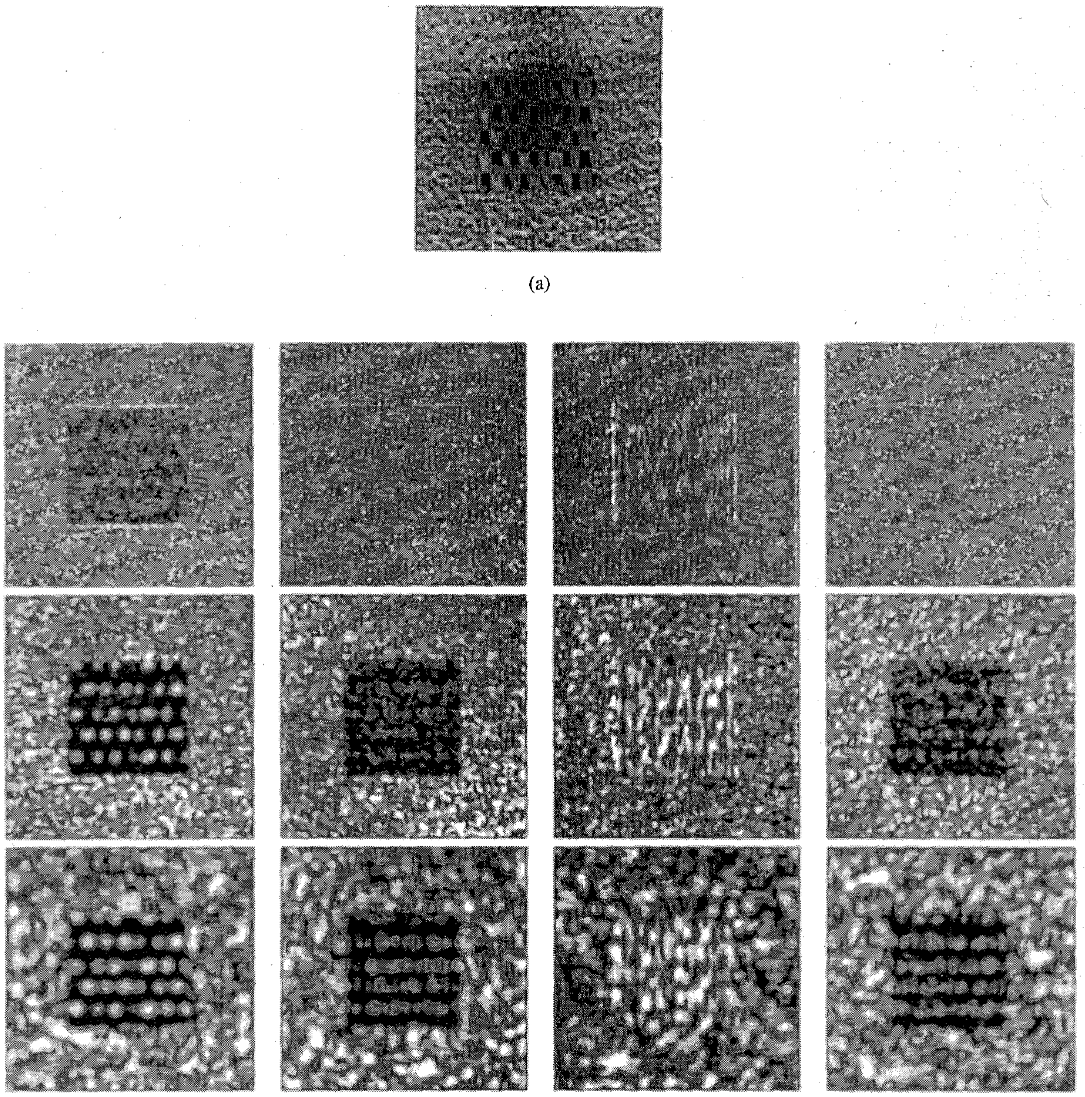


Fig. 5. Example of cover of the frequency plane (ω_x, ω_y) with dilated dyadic wavelets constructed by modulating a window $\theta(x, y)$ with sinusoidal waves having different orientations. Each circle symbolizes the frequency support of a dilated wavelet along a particular orientation.

The orientation of texture elements and their frequency contents seem to be important clues for discrimination. This has motivated early researchers to study the repartition of a texture energy in the Fourier domain. For segmentation purposes, it is however necessary to localize texture measurements over neighborhoods of varying sizes. It has thus been proposed to replace the Fourier transform with localized energy measurements at the output of filter banks that compute a wavelet-like transform [15], [29]. Besides the algorithmic efficiency of this approach, this model is partly supported by physiological studies of the visual cortex.

In the cat's visual cortex, Hubel and Wiesel [13] discovered a class of cells called simple cells, whose response depends upon the frequency and orientation of the visual stimuli. Numerous physiological experiments [28] have shown that their response can be modeled with linear filters, whose impulse response have been measured at different locations of the visual cortex. Daugmann [11] showed that these impulse response can be approximated by Gaussian windows modulated by a sinusoidal wave. Depending upon the cortical cell, this modulated Gaussian is dilated and has a specific spatial orientation tuning. These findings suggest the existence of some sort of wavelet transform in the visual cortex, combined with subsequent nonlinearities. The frequency resolution of these "physiological" wavelets seems to be of the order of 1–1.5 octaves.

Several texture discrimination algorithms [15], [29] are based on Gabor wavelets. These wavelets are constructed with a rotationally symmetric Gaussian window $\theta(x, y)$ that



(a)

(b)

Fig. 6. The original image is at the top. The energy of the wavelet transform $|T^k f(2^j, u, v)|^2$ is shown along four orientations at three scales 2^j, for $1 \leq j \leq 3$. From left to right, the orientation tuning of the wavelet are 0, 45, 90, and 135°, respectively. The scale increases from top to bottom. The energy value is large when the orientation and scale matches the texture structures. The distribution of wavelet energy across scales and orientations is different for the two textures.

is multiplied by sinusoidal waves that propagates along K orientations $\{\alpha_k\}_{1 \leq k \leq K}$, with different phases. Two wavelets with quadrature phases are the real and imaginary parts of a complex wavelet

$$\psi^k(x, y) = \theta(x, y) \exp\left[-i\xi(x \cos\alpha_k + y \sin\alpha_k)\right].$$

Since the wavelet has a frequency resolution of the order of one octave, we can restrict the scales to $\{2^j\}_{j \in Z}$ and define the wavelet transform in each direction α_k by

$$T^k f(2^j, u, v) = f \star \tilde{\psi}^k_{2^j}(u, v). \tag{10}$$

This wavelet representation is complete and stable if there exists $A > 0$ and B such that the Fourier transforms satisfy

$$\forall(\omega_x, \omega_y) \in \boldsymbol{R}^2, \quad A \leq \sum_{k=1}^{K} \sum_{j=-\infty}^{+\infty} |\hat{\psi}^k(2^j\omega_x, 2^j\omega_y)|^2 \leq B. \tag{11}$$

Let $\hat{\theta}(\omega_x, \omega_y)$ be the Fourier transform of $\theta(x, y)$. The Fourier transform of $\psi^k_{2^j}(x, y)$ is

$$\hat{\psi}^k_{2^j}(\omega_x, \omega_y) = \sqrt{2^j}\hat{\theta}(2^j\omega_x - \xi\cos\alpha_k, 2^j\omega_y - \xi\sin\alpha_k).$$

It is a translation and dilation of $\hat{\theta}(\omega_x, \omega_y)$. Its frequency energy is mostly concentrated around the frequency

$(\xi 2^j \cos\alpha_k, \xi 2^j \sin\alpha_k)$, in a neighborhood proportional to 2^{-j}. Fig. 5 shows an approximate frequency plane cover with such dyadic wavelets that satisfies (11). The filtering formula (10) shows that $|T^k f(2^j, u, v)|^2$ can be interpreted as a localized measurement of the frequency energy of the image f in the neighborhood of $(\xi \cos\alpha_k/2^j, \xi \sin\alpha_k/2^j)$. Varying the scale 2^j and the angle α_k modifies the frequency channel. Fig. 6 displays the wavelet transform energy along four orientations for a textured image. The energy $|T^k f(2^j, u, v)|^2$ is maximum when the angle α_k is along the orientation of the main texture components and when 2^j matches the scale of these structures. In Fig. 6, the center texture has its energy mostly concentrated in horizontal and vertical directions but has little energy in diagonal orientations. On the other hand, the texture at the periphery has no preferential orientation. The energy spread across scales is also different for these two textures. For segmentation, the main difficulty is to find an algorithm that aggregates the wavelet responses at all scales and orientations in order to find the boundaries of homogeneous textured regions. Indeed, within any single region, for each scale and orientation, the wavelet energy $|T^k f(2^j, u, v)|^2$ may have a relatively large degree of variability as shown by Fig. 6. Most algorithms attenuate these variations with a local spatial averaging of $|T^k f(2^j, u, v)|^2$. Clustering procedures [29], [15] or detection of sharp transitions over wavelet energy measurements [21] have been used to integrate the information across orientations and scales to produce a final segmentation. Despite their good experimental results, these algorithms are quite *ad hoc* and are probably not the final answer to texture discrimination. We also do not know precisely what are the classes of textures that are discriminated by these different segmentation procedures.

The formalization of texture recognition problems is often easier in a stochastic framework. It does not mean that textures are supposed to be created by some random physical processes, but we can model mathematically a class of images having the same texture as the realizations of a particular "texture process," i.e., wood. These texture processes are generally non-Gaussian and nonstationary and are thus particularly difficult to analyze. Moreover, when looking at a single texture, we see only one realization of this process. To identify a process from one realization is generally very difficult. Even if we suppose that the process is Gaussian, in which case it is characterized by its covariance, a reliable estimation of the covariance from a single realization is extremely hard when the process is not stationary. The Fourier basis diagonalizes the covariance matrix of stationary processes, and we thus only need to estimate the diagonal entries which correspond to the power spectrum. For more general nonstationary processes we do not know the basis which diagonalizes the covariance matrix. Understanding which class of bases are well adapted to estimate particular classes of texture processes is an open issue. These outstanding problems illustrate the difficulty to develop a coherent texture theory which is in accordance with human texture discrimination.

V. Conclusion

The scale is a fundamental parameter of visual processing, but wavelets are not the only tools to modify the scale and resolution of images. Other diffusion algorithms can remove certain image details and keep other components such as edges, by applying a nonlinear partial differential equation to the image. The application such multiscale nonlinear diffusions to edge detection has been introduced by Perona and Malik [27], and several important classes of nonlinear equations have been studied [2], [32]. Similar nonlinear diffusions are obtained by thresholding wavelet coefficients but no precise relations have been established between nonlinear PDE approaches and wavelet decompositions.

The wavelet mathematical theory is reaching a mature stage but how to use of this multiscale information for information processing is not always clear. The applications that we described illustrate the difficulty to design a non *ad hoc* interface between wavelet descriptors and classification algorithms. This leaves many opportunities for new research ideas in multiscale computer vision.

Acknowledgment

The author would like to thank F. Bergeaud and F. Falzon for help with the figures.

References

[1] J. Aloimonos and A. Rosenfeld, "Computer vision," *Sci.,* vol. 253, pp. 1249–1253, Sept. 1991.

[2] L. Alvarez, P.-L. Lions, and J.-M. Morel, "Image selective smoothing and edge detection by nonlinear diffusion II," *SIAM J. Numer. Anal.*, vol. 29, no. 3, pp. 845–866, 1992.

[3] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *Int. J. Computer Vision,* no. 2, pp. 283–310, 1989.

[4] Z. Berman and J. Baras, "Properties of multiscale maxima and zero-crossing representations," *IEEE Trans. Signal Process.,* vol. 41, pp. 3216–3231, Dec. 1993.

[5] A. Bovik, M. Clark, and W. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 12, pp. 55–73, 1990.

[6] P. Burt, "Smart sensing within a pyramid vision machine," *Proc. IEEE,* vol. 76, pp. 1006–1015, Aug. 1988.

[7] P. Burt and E. Adelson, "The Lapalacian pyramid as a compact image code," *IEEE Trans. Comm.*, vol. COM-31, pp. 532–549, Apr. 1983.

[8] J. Canny, "A computational approach to edge detection," *IEEE Trans. Patt. Anal. and Mach. Intell.,* vol. 36, pp. 961–1005, Sept. 1986.

[9] B. Horn, *Robot Vision.* Cambridge, MA: MIT Press, 1985.

[10] A. Cohen and J. Kovačević, "Wavelets: The mathematical background," *Proc. IEEE,* this issue, pp. xxx–xxx.

[11] J. G. Daugmann, "Two-dimensional spectral analysis of cortical receptive field profile," *Vision Res.*, vol. 20, pp. 847–856, 1980.

[12] W. Grimson, "Computational experiments with a feature based stereo algorithm," *IEEE Trans. Patt. Anal. and Mach. Intell.,* vol. 7, pp. 17–34, Jan. 1985.

[13] D. Hubel and T. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, 1962.

[14] R. Hummel and R. Moniot, "Reconstruction from zero-crossings in scale-space," *IEEE Trans. Acoust. Speech and Signal Process.,* vol. 37, Dec. 1989.

[15] A. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters,"*Patt. Recog.,* vol. 24, no. 12, pp. *1167–1186, 1991.*

[16] B. Julesz, "Textons, the elements of texture perception and their interactions," *Nature*, vol. 290, Mar. 1981.

[17] G. Kanizsa, *Organization in Vision.* New York: Praeger, 1979.

[18] J. Koenderink, "The structure of images," in *Biological Cybernetics.* New York: Springer Verlag, 1985.

[19] K. Kumaran, D. Geiger, and L. Gurvits, "Illusory surfaces and visual organization," *J. Network: Computat. in Neur. Syst.*, vol. 7, no. 1, Feb. 1995.

[20] A. Laine and J. Fan, "Texture discrimination and classification by wavelet packets," *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 15, pp. 1186–1191, Nov. 1993.

[21] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *J. Opt. Soc. Amer.*, vol. 7, pp. 923–932, 1990.

[22] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 11, pp. 674–693, July 1989.

[23] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 14, pp. 710–732, July 1992.

[24] D. Marr, *Vision.* Freeman: San Fransisco, 1982.

[25] Y. Meyer, *Ondelettes et Operateurs.* Paris: Hermann, 1990.

[26] ——, *Wavelets, Algorithms and Applications.* Philadelphia: SIAM, 1993.

[27] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 12, pp. 629–639, July 1990.

[28] D. A. Pollen and S. F. Ronner, "Visual cortical neurons as localized spatial frequency filter," *IEEE Trans. Syst. Man. Cybern.*, vol. 13, Sept. 1983.

[29] M. Porat and Y. Zeevi, "Localized texture processing in vision: analysis and synthesis in Gaborian space," *IEEE Trans. Biomed. Eng.*, vol. 36, pp. 115–129, Jan. 1989.

[30] A. Rosenfeld and M. Thurston, "Edge and curve detection for visual scene analysis," *IEEE Trans. Comput.*, vol. C-29, 1971.

[31] A. Rosenfeld and G. J. Vanderburg, "Coarse-fine template matching," *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-7, pp. 104–107, 1977.

[32] G. Sapiro and A. Tannenbaum, "On affine plane curve evolution," *J. Function. Anal.*, vol. 119, pp. 514–529, 1994.

[33] E. Schwartz, "Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding," *Vision Res.*, vol. 20, pp. 645, 1980.

[34] M. Unser and M. Eden, "Multiresolution feature extraction and selection for texture feature segmentation," *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. 11, pp. 717–727, 1989.

[35] Z. Cvetkovic and M. Vetterli, "Discrete-time wavelet extrema representation: design and consistent reconstruction," to be published.

[36] A. Witkin, "Scale space filtering," in *Proc. Int. Joint. Conf. Artif. Intell.*, 1983.

[37] A. Yuille and T. Poggio, "Scaling theorems for zero-crossings," *IEEE Trans. Patt. Anal. and Mach. Intell.*, vol. PAMI-8, Jan. 1986.

Stéphane Mallat graduated from Ecole Polytechnique, Paris, and from the Ecole Nationale Superieure des Télécommunications, Paris, in 1984 and 1985, respectively. He received the Ph.D. degree in electrical engineering from the University of Pennsylvania, Philadelphia, PA, in 1988.

In 1988, he joined the Courant Institute of Mathematical Sciences at New York University, New York, NY. He is currently Associate Professor of Computer Science at the Courant Institute and Professor in the Applied Mathematics Departement of Ecole Polytechnique, Paris, France. His research interests include computer vision, signal processing and applied mathematics.

Dr. Mallat received the 1990 IEEE Signal Processing Society's paper award and the Alfred Sloan Fellowship in Mathematics, in 1993.