Depth Estimation from Image Structure

Antonio Torralba and Aude Oliva

Abstract—In the absence of cues for absolute depth measurements as binocular disparity, motion, or defocus, the absolute distance between the observer and a scene cannot be measured. The interpretation of shading, edges, and junctions may provide a 3D model of the scene but it will not provide information about the actual "scale" of the space. One possible source of information for absolute depth estimation is the image size of known objects. However, object recognition, under unconstrained conditions, remains difficult and unreliable for current computational approaches. Here, we propose a source of information for absolute depth estimation based on the whole scene structure that does not rely on specific objects. We demonstrate that, by recognizing the properties of the structures present in the image, we can infer the scale of the scene and, therefore, its absolute mean depth. We illustrate the interest in computing the mean depth of the scene with application to scene recognition and object detection.

Index Terms—Depth, image statistics, scene structure, scene recognition, scale selection, monocular vision.

1 Introduction

The fundamental problem of depth perception from monocular information is illustrated in Fig. 1. In the absence of cues for absolute depth measurement, such as binocular disparity, motion, or defocus, the three cubes will produce the same retinal image and, therefore, the absolute distance between the observer and each cube cannot be measured. The interpretation of shading, edges and junctions may provide a 3D model of the cube (relative depth between parts of the cube) but it will not inform about its actual size. This ambiguity problem does not apply however when dealing with real-world stimuli (Fig. 1b). Physical processes that shape natural structures are different at each scale (e.g., leaves, forests, mountains). Humans also build different types of structures at different scales, mainly due to functional constraints in relation with human size (e.g., chair, building, city). As a result, different laws with respect to the building blocks, the way that they are organized in space and the shape of the support surfaces, govern each spatial scale [14].

The constraints on the structure of the 3D scene at each spatial scale can be directly transposed into image content. Fig. 2 shows three pictures representing environments with different mean depths: The scenes strongly differ in their global configuration, the size of the component surfaces, and the types of textures. Specifically, panoramic views typically display uniform texture zones distributed along horizontal layers. Views of urban environments in the range of a few hundred meters show dominant long horizontal and vertical contours and complex squared patterns. Close-up views of objects tend to have large flat surfaces and, on average, no clear dominant orientations [22]. As the observed scale directly depends on the depth of the view, by *recognizing* the properties of the image structure, we can infer the scale of the scene and, therefore, the absolute depth.

Most of the techniques for recovering depth information focus on relative depth information: shape from shading [13], from texture gradients [35], from edges and junctions [2], from symmetric patterns [32], from Fractal dimension [16], [26] and from other pictorial cues such as occlusions, relative size, and elevation with respect the horizon line [24]. Most of these techniques apply only to a limited set of scenes. The literature on absolute depth estimation is also very large but the proposed methods rely on a limited number of sources of information (e.g., binocular vision, motion parallax, and defocus). However, under normal viewing conditions, observers can provide a rough estimate of the absolute depth of a scene even in the absence of all these sources of information (e.g., when looking at a photograph). One additional source of information for absolute depth estimation is the use of the size of familiar objects like faces, bodies, cars, etc. However, this strategy requires decomposing the image into its constituent elements. The process of image segmentation and object recognition, under unconstrained conditions, remains still difficult and unreliable for current computational approaches. The method proposed in this paper introduces a source of information for absolute depth computation from monocular views that does not require parsing the image into regions or objects: the global image structure. The underlying hypothesis of this approach is that the recognition of the scene as a whole is a simpler problem than the one of general object detection and recognition [21], [22], [37].

2 Image Structure

In recent years, there have been an increasing number of models of the image structure based on simple image statistics (e.g., [4]). These models have been motivated by applications in image indexing and computation of similarities between pictures (e.g., [3], [5], [7], [10], [20], [39]) and

- *A. Torralba is with the Department of Brain and Cognitive Sciences, E25-201, MIT, 45 Carleton Street, Cambridge, MA 02139. E-mail: torralba@ai.mit.edu.*
- *A. Oliva is at the Center for Ophthalmic Research, Brigham and Women's Hospital, 221 Longwood Avenue, Boston, MA 02115. E-mail: oliva@search.bwh.harvard.edu.*

Manuscript received 23 Feb. 2001; revised 28 Feb. 2002; accepted 1 Mar. 2002. Recommended for acceptance by R. Sharma. For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 113671.

0162-8828/02/$17.00 © 2002 IEEE

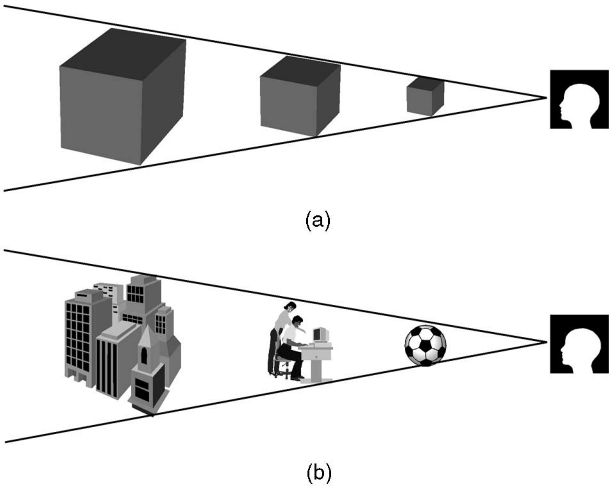


Fig. 1. (a) Artificial stimulus: The monocular information cannot provide an absolute depth percept. (b) Real-world stimulus: The recognition of image structures provides unambiguous monocular information about the absolute depth between the observer and the scene.

the study of models for natural images (e.g., [8], [17], [23], [30]). For the purpose of this section, we consider a simple definition of the image structure based on a description of the textural patterns present in the image and their spatial arrangement [21], [22], [37]. In this section, we discuss two levels of description of the image structure based on the second order statistics of images (Fig. 2). The first level, the magnitude of the global Fourier transform of the image, contains only unlocalized information about the dominant orientations and scales that compose the image. The second level, the magnitude of a local wavelet transform, provides the dominant orientations and scales in the image with a coarse description of their spatial distribution.

2.1 Unlocalized Image Structure and Spectral Signatures

The discrete Fourier transform (FT) of an image is defined as:

$$I(\mathbf{f}) = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} i(\mathbf{x})\, h(\mathbf{x})\, e^{-j2\pi<\mathbf{f},\mathbf{x}>}, \qquad (1)$$

where $i(\mathbf{x})$ is the intensity distribution of the image along the spatial variables $\mathbf{x} = (x, y)$; $j = \sqrt{-1}$; and the spatial frequency variables are defined by $\mathbf{f} = (f_x, f_y) \in [-0.5, 0.5] \times [-0.5, 0.5]$ (units are in cycles per pixel); and $h(\mathbf{x})$ is a circular window that reduces boundary effects. The amplitude spectrum is defined as the magnitude of the FT: $A(\mathbf{f}) = |I(\mathbf{f})|$. The amplitude spectrum reveals the dominant orientations and textural patterns in the image (Fig. 2). It is acknowledged that the information concerning spatial arrangements and shapes of the structures in the image are contained in the phase function of the FT. In fact, if we consider an image as being any possible distribution of pixel intensities, then the amplitude spectrum is not informative because many very different images would have the same amplitude spectrum. However, in the context of real-world scene pictures, the amplitude spectrum has a strong relationship with the spatial structure of the scene [22]. In order to study the relationship between the image amplitude spectrum and the scene structure, we define the *spectral signature* of a set of images S as the mean amplitude spectrum:

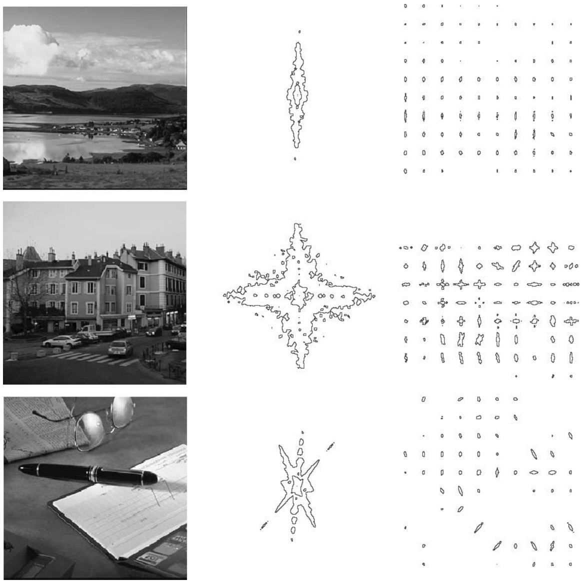


Fig. 2. Three examples of images used in this study. The scenes strongly differ in their absolute mean depth. For each scene, we show the sections of the global magnitude of the Fourier transform (center) and the sections of the magnitude of the windowed Fourier transform (right).

$$\overline{A}_S(\mathbf{f}) = E[A(\mathbf{f}) \mid S], \qquad (2)$$

where E is the expectation operator. The spectral signature $\overline{A}_S(\mathbf{f})$ reveals the dominant structures shared by the images of the set S. Several studies (e.g., [8], [30]) have observed that the averaged amplitude spectrum of the set of real-world scene pictures falls with a form: $\overline{A}_S \sim 1/\|\mathbf{f}\|^{\alpha}$ with $\alpha \sim 1$. Real-world scenes can be divided into semantic categories that depict specific spectral signatures (see [22] for a detailed discussion). The clearest example of picture sets distinguished by their spectral signatures is man-made versus natural structures. Both spectral signatures are defined by the conditional expectations:

$$\overline{A}_{art}(\mathbf{f}) = E[A(\mathbf{f}) \mid \text{man-made}] \qquad (3)$$

$$\overline{A}_{nat}(\mathbf{f}) = E[A(\mathbf{f}) \mid \text{natural}]. \qquad (4)$$

Fig. 3 shows the contour plots of the spectral signatures obtained from more than 6,000 pictures (see a description of the database in Section 6). $\overline{A}_{art}(\mathbf{f})$ has dominant horizontal and vertical orientations due to the bias found in man-made structures [1], [17], [22]. $\overline{A}_{nat}(\mathbf{f})$ contains energy in all orientations with a slight bias toward the horizontal and the vertical directions. These spectral characteristics are shared by most of the pictures of both categories allowing the discrimination between man-made and natural scenes with a very high confidence (93.5 percent, refer to Section 5.1 and [22], [37]).

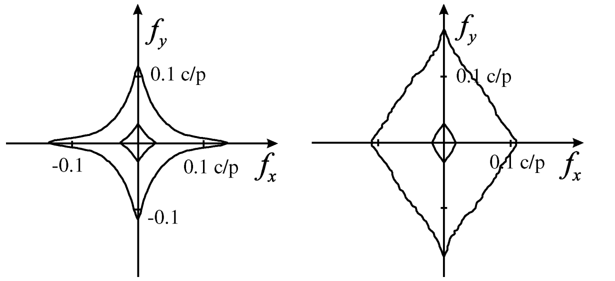


Fig. 3. Global spectral signatures of man-made and natural environments averaged from more than 6,000 images. The contour plots represent the 50 percent and the 80 percent of the energy of the spectral signatures. Units are in cycles per pixel.

2.2 Spatial Organization and Local Spectral Signatures

An essential aspect of an image representation, that the global amplitude spectrum does not encode, concerns the spatial arrangement of the structural elements in the image [3], [5], [7], [22], [37]. For example, panoramic landscapes have the sky at the top, characterized by low spatial frequencies, the horizon line around the center, and usually texture at the bottom part. Urban scenes in the range of a few hundred meters will have the sky at the top, buildings in the middle part, and a road at the bottom. That specific arrangement produces a particular spatial pattern of dominant orientations and scales (Fig. 2) that can be described using a wavelet transform:

$$I(\mathbf{x},k) = \sum_{x'} i(\mathbf{x}')\; h_k(\mathbf{x}-\mathbf{x}'). \qquad (5)$$

Different choices of the functions h_k provide different representations (e.g., [33]). One of the most common functions are complex Gabor wavelets: $h_k(\mathbf{x}) = e^{-\pi\|\mathbf{x}\|^2/\sigma_k^2} e^{-2\pi j<\mathbf{f}_k,\mathbf{x}>}$. In such a representation, $I(\mathbf{x},k)$ is the output at the location $\mathbf{x}$ of a complex Gabor filter tuned to the spatial frequency defined by $\mathbf{f}_k$. The resulting representation encodes local scale and orientation information. When $h_k(\mathbf{x}) = h_r(\mathbf{x})\; e^{-j\,2\pi<\mathbf{f}_k,\mathbf{x}>}$ ($h_r(\mathbf{x})$ is a window with a circular support of constant radius r), then $I(\mathbf{x},k)$ corresponds to the Windowed Fourier transform (WFT) and can be more conveniently written as $I(\mathbf{x},\mathbf{f})$ with $\mathbf{f}$ defined as in (1).

The magnitude $A(\mathbf{x},k) = |I(\mathbf{x},k)|$ provides local structural information in a neighborhood of the location $\mathbf{x}$. Both the Gabor transform and the WFT provide similar structural information. For visualization purposes, we will use the WFT for the figures in this paper: $A(\mathbf{x},\mathbf{f}) = |I(\mathbf{x},\mathbf{f})|$. Fig. 2 shows sections of the WFT at 10 × 10 locations.

The local spectral signature of a set of images S is defined as follows:

$$\overline{A}_S(\mathbf{x},k) = E[A(\mathbf{x},k)\,|\,S]. \qquad (6)$$

The local spectral signature of a set of images gives information about the dominant spectral features and their mean spatial distribution. Fig. 5 shows examples of local spectral signatures (WFT with $r = 16$ pixels) for man-made and natural scene pictures with different depths (see next section) and illustrates that even this simple statistics are nonstationary when considering specific sets of images (e.g., man-made environments in the range of 1Km).

3 Image Structure as a Depth Cue

In this paper, we refer to the mean depth of a scene image as a measure of the scale or size of the space that the scene subtends. This section presents the main sources of variability found in the spectral features with respect to the mean depth.

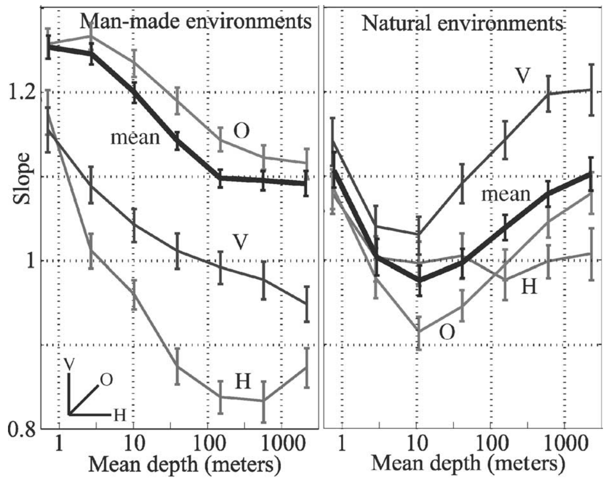
A number of studies have focused in the study of the scale invariance property of natural image statistics (e.g. [8], [17], [28]). Most of these studies focus on small scaling changes (as the ones that occur within a single image) and do not use absolute units for the scale (images depicting structures at different sizes in the real world are averaged together). However, when considering large scale changes (a scaling factor larger than 10), there exist significant differences between the statistics of pictures depicting scenes and objects at different scales in absolute units (e.g., meters). There are at least two reasons that can explain the dependency between the image structure and the scene mean depth:

- The *point of view*: Under normal viewing conditions, the point of view that an observer adopts on a specific scene is strongly constrained. Objects can be observed under almost any point of view. However, as distance and scale overtake human scale, the possible viewpoints become more limited and predictable [6]. The dominant orientations of the image strongly vary with the point of view (e.g., vanishing lines, [6], [22]), and, consequently, the spatial arrangement of the main structures (e.g., position of the ground level, horizon line).
- The *building blocks*: The building blocks (or primitives [17]) refer to the elements (surfaces and objects) that compose the scene. The building blocks (their shape, texture, color, etc.) largely differ between natural and man-made environments, as well as between indoor and outdoor places [22], [36], [39]. The building blocks of a scene also differ strongly from one spatial scale to another due to functional constraints and to the physical processes that shape the space at each scale.

Both the building blocks and the point of view of the observer determine the dominant scales and orientations found in an image. In the following, we discuss the relationship between the image spectral components (global and local) and the mean depth of the scene.

3.1 Relationship between the Global Spectral Signature and Mean Depth

For the range of distances that we are working with (from centimeters to kilometers), the problem of scaling cannot be modeled by a zoom factor with respect to one reference image. As the image is limited in size and resolution, by zooming out the image by a factor larger than 2, new structures appear at the boundaries of the image and, because of the sampling, small details disappear. The resulting new picture gets a completely different spatial shape and a new amplitude spectrum that cannot be related to the one of the original image by a simple scaling operation. In order to study the variations of scene structure for different depths, we use the spectral signatures (2). It is



(a) (b)

Fig. 4. Evolution of the slope of the global magnitude spectrum of real-world pictures with respect to the mean depth of the scene. The picture shows the evolution of the slope $\alpha_D(\theta)$ at 0 (Horizontal; f_x), 45 (Oblique) and 90 (Vertical; f_y) degrees, and its mean value averaged for all the orientations. Error bars represent 90 percent intervals obtained by Bootstrap.

interesting to make the distinction between man-made and natural structures as they strongly differ in the building blocks and have different relationships between structure and spatial scale or distance. Considering first man-made structures, we define S as the set of pictures of man-made scenes sharing the same mean distance (D) from the observer. The spectral signature is:

$$\overline{A}_{D,art}(\mathbf{f}) = E[A(f)\,|\,D,\ \text{man-made}]. \qquad (7)$$

Fig. 5 shows the spectral signatures for different ranges of depth. The spectral signatures can be modeled by: $\overline{A}_D(\mathbf{f}) \sim \Gamma_D(\theta)/\|\mathbf{f}\|^{\alpha_D(\theta)}$, as proposed in [22], where $\theta = \angle\mathbf{f}$. $\Gamma_D(\theta)$ is a magnitude prefactor that is a function of orientation. The spectral signature has a linear decaying rate in logarithmic units with a slope given by $-\alpha_D(\theta)$ ([22], [30]). These two functions can be obtained by a linear fitting at each orientation of the spectral signature in logarithmic units [30]. Fig. 4 shows the mean slope $\overline{\alpha}$ (averaged with respect to orientation), for different depths. The mean slope reveals the mean fractal dimensionality of the picture, which is related to its apparent roughness [26] or clutter [17]. An increasing of the slope means a reduction of energy in the high spatial frequencies, which thus changes the apparent roughness of the picture. For man-made structures (Fig. 4a), we observe a monotonic decreasing slope (i.e., increasing roughness) when increasing depth. This is an expected result as close-up views on man-made objects contain, on average, large flat surfaces and homogeneous regions (i.e., low roughness). When increasing the distance, surfaces are likely to break down in small pieces (objects, walls, doors, windows, etc.) increasing, therefore, the global roughness of the picture (Fig. 5).

Although the increase of clutter with distance appears as something intuitive, it is not a general rule and it cannot be applied to every picture. For natural structures, the spectral signature exhibits a completely opposite behavior with respect to the mean depth (see Figs. 4b and 5): the mean slope increases when depth increases. This fact is related to a decreasing of the mean roughness of the picture, with distances. Close-up views on natural surfaces are usually textured, giving to the amplitude spectrum a small decaying slope. When distance increases, natural structures become larger and smoother (the small grain disappears due to the spatial sampling of the image). The examples in Fig. 5 illustrate this point. For natural scene pictures, on average, the more we increase the mean depth of the scene the more energy concentrates in the low spatial frequencies, which is the opposite behavior with respect to man-made structures.

The dominant orientations also provide relevant depth related information (Fig. 5). To illustrate this point, Fig. 4 shows the slopes for the horizontal, oblique, and vertical spectral components for both man-made and natural structures at different scales. For instance, many panoramic views have a straight vertical shape in their amplitude spectrum due to the horizon line. City-center views have similar quantities of horizontal and vertical orientations and only a little energy for the oblique orientations. On average, close-up views of objects have no strong dominant orientations and, thus, an isotropic amplitude spectrum.

3.2 Relationship between the Local Spectral Signatures and Depth

As in (7), we can study the averaged local amplitude spectrum (local spectral signatures) for different depths. Fig. 5 shows the evolution of the local spectral signatures with respect to depth for man-made and natural structures. We can see that not only the general aspect of the local spectral signatures changes with depths but also the spatial configuration of orientation and scales. Note that the variations are mostly from top to bottom. The typical behavior can be summarized as follows:

- An increase of the global roughness with respect to depth for man-made structures and a decrease of global roughness for natural structures.
- For near distances ($D < 10$m), the spectral signatures are stationary and there is almost no bias towards horizontal and vertical orientations.
- For intermediate distances (10m to 500m) the spectral signatures are nonstationary and biased towards horizontal and vertical orientations. On average, the scene structure is dominated by smooth surfaces on the bottom (e.g., support surfaces like roads, lakes, fields) and also on the top due to the sky. The center contains buildings with high frequency textures with cross-like spectra for man-made environments or almost isotropic textures for natural environments.
- For far distances (> 500m), the sky introduces a smooth texture in the top part. A long horizontal plane, filled with small squared man-made structures or with a smooth natural texture, usually dominates the bottom zone.

To summarize, there exists a strong relationship between the structures present in the image and the mean depth of the scene. This point is demonstrated in the rest of the paper by showing that absolute depth can be estimated from structural features.

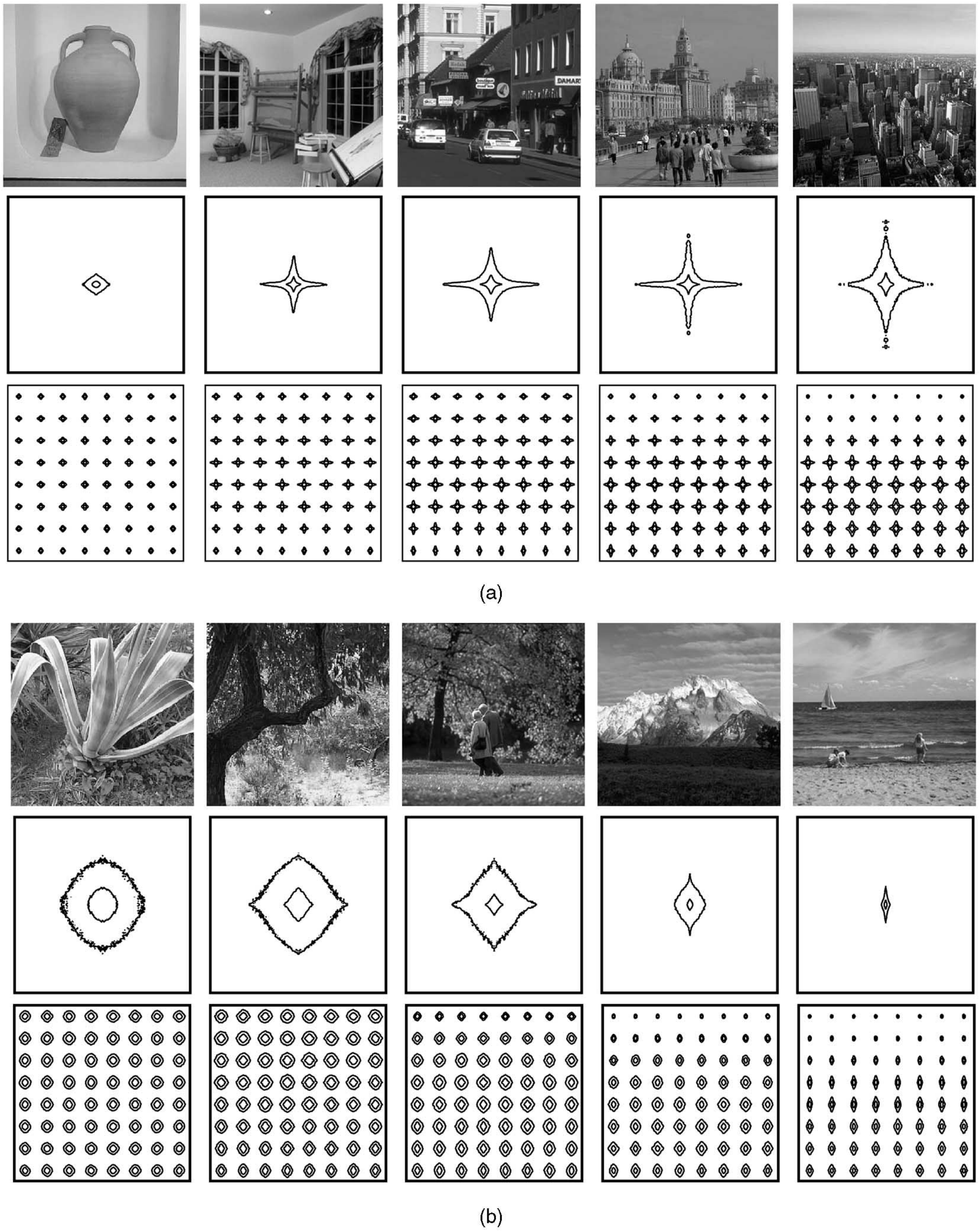


Fig. 5. Evolution of the global and local spectral signatures of (a) man-made and (b) natural scenes with respect to the mean depth. Each signature is obtained from averaging over more than 300 pictures with depths of, from left to right, 1, 10, 100, 1,000 meters and panoramic views ($D > 1$ Km).

4 Low-Dimensional Representations of Image Structure

The image features given by $A(\mathbf{f})$ and $A(\mathbf{x}, k)$ (Section 2), are very high-dimensional. In this section, we discuss low-dimensional representations of the image structure based on those features and we review other structural representations based on higher order statistics.

A number of low-dimensional representations based on the statistics of wavelet coefficients (tuned to different orientations and scales) have been used with success in texture [12], [27], object [31] and scene representations [10], [22], [39]. Representations based on global statistics are suited for the description of textures [12], [27], [40]. Global statistics assume stationarity and are computed by averaging measures across the entire image. Although they are not appropriate when representing nonstationary images, like pictures of specific objects or scene categories, global statistics also provide useful information for recognition tasks (e.g., [22]).

As discussed in Section 2, one of the simpler global statistics is the output energy of each wavelet (5):

$$A_k^2 = \sum_{\mathbf{x}} |I(\mathbf{x}, k)|^2 = \int A(\mathbf{f})^2 H_k(\mathbf{f})\, d\mathbf{f}. \qquad (8)$$

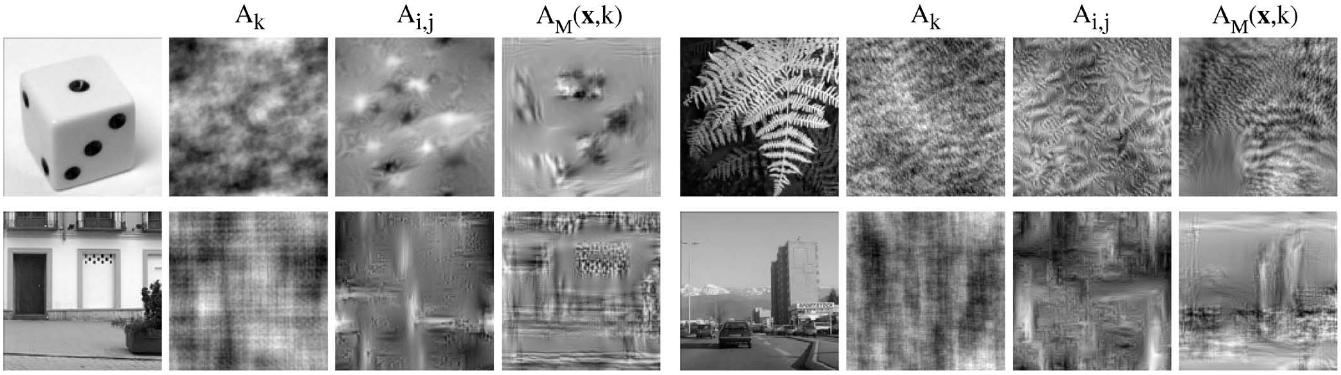


Fig. 6. Comparison of the information available in the representations given by the global output energy of the wavelets, A_k, the cross-correlation of wavelet coefficients, $A_{i,j}$, and the local output energy of the analyzing wavelets, $A_M(\mathbf{x},k)$. The three representations use Gabor wavelets tuned to 4 scales and 4 orientations per scale. For each picture, we show three texture-like images obtained by coercing noise to have the same features than the original image.

The features A_k^2 are samples of the power spectrum of the image, $A(\mathbf{f})^2$, averaged with by the FT transform, $H_k(\mathbf{f})$, of the wavelet $h_k(\mathbf{x})$. A_k encodes information about the second order statistics of the image. K is the number of wavelets used in the decomposition and, therefore, the dimensionality of the representation. This representation encodes the dominant orientations and scales in the image. There is no information about the spatial organization or the shape of the objects in the image. In order to illustrate the information preserved by the representation, Fig. 6 shows examples of real-world pictures and synthesized texture-like images that have the same values of A_k. Therefore, the real-world images and the synthesized textures are indistinguishable by this representation.

According to studies of the statistics of natural images (e.g., [17], [27], [34]), higher order statistics show also particular characteristics for real-world pictures. For instance, the magnitude of the wavelet coefficients at different orientations and scales are correlated. The correlations coefficients are a function of the elements that compose the image and their spatial distribution (local sharp edges give rise to high correlation coefficients [27]). The magnitude correlations can be written as:

$$A_{i,j}^2 = \sum_{\mathbf{x}} |I(\mathbf{x},i)||I(\mathbf{x},j)|. \tag{9}$$

Note from (8), that this representation includes also $A_k = A_{k,k}$. The dimensionality of this representation is K^2, although it can be reduced by computing only the correlations of adjacent wavelet coefficients (in terms of scale and orientation, [27]). Fig. 6 shows examples of synthesized texture-like images that have the same magnitude correlations $A_{i,j}$ than the original picture. The generated textures preserve additional information about the original image. Apart from the orientations and scales, the magnitude correlations also preserve basic information about the sparseness and degree of clutter of edges and shapes found in the original picture.

More complete representations have been proposed in the field of texture synthesis [12], [27], [40]. The marginal histograms of the wavelet coefficients have been extensively used for texture analysis and synthesis (e.g., [12]). The dimensionality of the representation is BK where B is the number of bins (used for the histogram estimation) and K is the number of wavelets. The joint histogram of wavelet coefficients [31] gives a complete representation of the distribution but with much higher dimensionality (B^K) than the previous ones.

Spatial organization, although not relevant for texture representations, is a key attribute in scene representations (e.g., [5], [22]). Spatial information can be introduced by computing local statistics, averaging measures over local neighborhoods at different spatial locations. This provides a way of dealing with the nonstationary behavior of real images. The local energy of the wavelet coefficients is:

$$A_M^2(\mathbf{x},k) = \left\{ |I(\mathbf{x},k)|^2 \downarrow M \right\}. \tag{10}$$

For simplicity, the notation $\downarrow M$ represents the operation of downsampling in the spatial domain until the resulting representation $A_M(\mathbf{x},k)$ has a spatial resolution of M^2 pixels. Therefore, $A_M(\mathbf{x},k)$ has a dimensionality M^2K. Note that the global statistics ((8)) are obtained when $M=1$: $A_k = A_{M=1}(\mathbf{x},k)$. Fig. 6 shows synthesized texture-like images that are constrained to have the same features $A_M(\mathbf{x},k)$, with $M=8$, than the original image. Similarly, we can define local higher-order statistics as: $A_M^2(\mathbf{x},i,j) = \{|I(\mathbf{x},i)I(\mathbf{x},j)| \downarrow M\}$.

Each scene picture is represented by a features vector $\mathbf{v}$ that contains the set of statistical measurements. Applying simple techniques such as PCA, we can further reduce the dimensionality of the features vector while preserving most of the information that accounts for the variability among different real-world pictures. The principal components PCs are the eigenvectors of the covariance matrix $\mathbf{C} = E[(\mathbf{v}-\mathbf{m})(\mathbf{v}-\mathbf{m})^T]$, where $\mathbf{v}$ is a column vector composed by the image features, and $\mathbf{m} = E[\mathbf{v}]$. Expectation is approximated by the average with respect to the entire image database.

In the rest of the paper, we refer to the column vector $\mathbf{v}$ as the L-dimensional vector obtained by projection of the image features onto the first L PCs with the largest eigenvalues.

In Section 6, we study the performance of the global (energy and magnitude correlation) and the local image statistics in predicting the mean depth of the scene. First, we introduce the learning framework used for modeling the relationship between scene mean depth and image statistics.

5 Probabilistic Model

In contrast to computational studies dedicated to depth perception based on predefined laws (stereo disparity, motion parallax, defocus, shading, texture gradients, etc.), the system we introduce in this section is designed to learn the relationship between the structures present in the picture and the mean depth of the scene. As discussed in Section 3, the relationship between structure and depth comes from the particular way that the world appears (is built) at each scale. For instance, the system has to learn that long oblique contours in a natural landscape scene are likely to correspond to a very large-scale structure (e.g., a mountain) and that the texture introduced by trees belongs to a medium-scale structure.

5.1 Depth Estimation

Our objective is to estimate the absolute mean depth of the scene, D, by means of the image features, $\mathbf{v}$. The function that minimizes the mean square error between the estimated and the real depth is the conditional expected value (see [25]):

$$\hat{D} = E[D \mid \mathbf{v}] = \int_{-\infty}^{\infty} D\, f_{D|v}(D \mid \mathbf{v})\, dD, \tag{11}$$

where

$$f_{D|v}(D \mid \mathbf{v}) = f_{D,v}(D, \mathbf{v}) / f_v(\mathbf{v})$$

and $f_v(\mathbf{v}) = \int f_{D,v}(D, \mathbf{v})\, dD$. The joint probability density function (PDF) $f_{D,v}(D, \mathbf{v})$ characterizes the relationship between the two random variables D and $\mathbf{v}$. As shown before, the relationship between the depth and the image statistics strongly differ between man-made and natural scenes. For this reason, the image database was split in two complementary groups: man-made (art) and natural (nat) scenes. Note that both groups may contain images with both natural and man-made structures, such as trees in a street or a farm in a field.

As shown in Figs. 3, 5, and 6, there are strong differences in the spectral characteristics of man-made and natural scenes. Therefore, we can expect to have high confidence discrimination even when using only unlocalized structural information, (8) [22]. Discrimination between man-made and natural structures can be done by computing the conditional probabilities for one image to belong to the art group or to the nat group, once the image features have been measured. One scene is considered as belonging to the art group if $f(\mathbf{v} \mid art) > f(\mathbf{v} \mid nat)$, with $p(art) = p(nat)$. The learning of the functions $f(\mathbf{v} \mid art)$ and $f(\mathbf{v} \mid nat)$ is performed using a mixture of Gaussians and the EM algorithm. We trained a classifier using the global structural features (8) and 2,000 pictures (1,000 for each category). One Gaussian cluster was enough for modeling the PDFs. The test was performed on 2,000 new pictures per group. The classification rate between man-made and natural structures was 93.5 percent. Other authors using other features have obtained similar results [10], [39]. For the rest of the paper, we study performances in depth estimation separately for both man-made and natural structures.

5.2 Learning

For building the depth estimator, we need to estimate the joint PDFs $f(D, \mathbf{v} \mid art)$ and $f(D, \mathbf{v} \mid nat)$. In the framework of regression algorithms, several approaches have been proposed for the estimation of the joint PDF [9]. Here, we used cluster-weighted modeling ([9], p. 178) as it provides a simple algorithm for the learning stage. For completeness, we reproduce here the main expressions of the estimation algorithm. In such a model, the joint PDF is expanded as a sum of Gaussian clusters, each one modeling locally the relationship between the input and the output distributions:

$$f(D, \mathbf{v} \mid art) = \sum_{i=1}^{N_c} g(D \mid \mathbf{v}, c_i)\, g(\mathbf{v} \mid c_i),\, p(c_i), \tag{12}$$

where D refers to depth and $\mathbf{v}$ to the image features. N_c corresponds to the number of clusters used for the approximation. Each cluster is decomposed in three factors: $p(c_i)$ is the weight of each cluster, $g(\mathbf{v} \mid c_i)$ is a multivariate Gaussian, with mean μ_i and covariance matrix $\mathbf{X}_i$, that defines the domain of influence in the input space of the cluster:

$$g(\mathbf{v} \mid c_i) = \frac{\exp\left[-\frac{1}{2}(\mathbf{v} - \mu_i)^T \mathbf{X}_i^{-1} (\mathbf{v} - \mu_i)\right]}{(2\pi)^{L/2} |\mathbf{X}_i|^{1/2}} \tag{13}$$

and $g(D \mid \mathbf{v},\, c_i)$ models the output distribution of the cluster:

$$g(D \mid \mathbf{v}, c_i) = \frac{\exp\left[-\left(D - a_i - \mathbf{v}^T \vec{b}_i\right)^2 / 2\sigma_i^2\right]}{\sqrt{2\pi}\sigma_i}. \tag{14}$$

This distribution is a Gaussian function with respect to D, with variance σ_i^2 and a mean that has a linear dependence on the image features: $a_i + \mathbf{v}^T \mathbf{b}_i$. T denotes the transpose. The model parameters, $p(c_i)$, μ_i, $\mathbf{X}_i$, σ_i^2, a_i, and $\mathbf{b}_i$, with $i = 1, \ldots, N_c$, are estimated using the EM algorithm [9], [15]. Let $\{D_t\}_{t=1,\ldots,N_t}$ and $\{\mathbf{v}_t\}_{t=1,\ldots,N_t}$ be the training data set (D_t are the depths, in logarithmic units, of a set of pictures and v_t are their respective structural features). The EM algorithm is an iterative procedure composed of two steps:

- E-step: Computes the posterior probabilities of the clusters given the observed data. For the k iteration:

$$p^k(c_i \mid D_t, \mathbf{v}_t) = \frac{g^k(D_t \mid \mathbf{v}_t, c_i)\, g^k(\mathbf{v}_t \mid c_i)\, p^k(c_i)}{\sum_{i=1}^{N_c} g^k(D_t \mid \mathbf{v}_t, c_i)\, g^k(\mathbf{v}_t \mid c_i)\, p^k(c_i)}. \tag{15}$$

- M-step: Computes the most likely cluster parameters:

$$p^{k+1}(c_i) = \frac{\sum_{t=1}^{N_t} p^k(c_i \mid D_t, \mathbf{v}_t)}{\sum_{i=1}^{N_c} \sum_{t=1}^{N_t} p^k(c_i \mid D_t, \mathbf{v}_t)}. \tag{16}$$

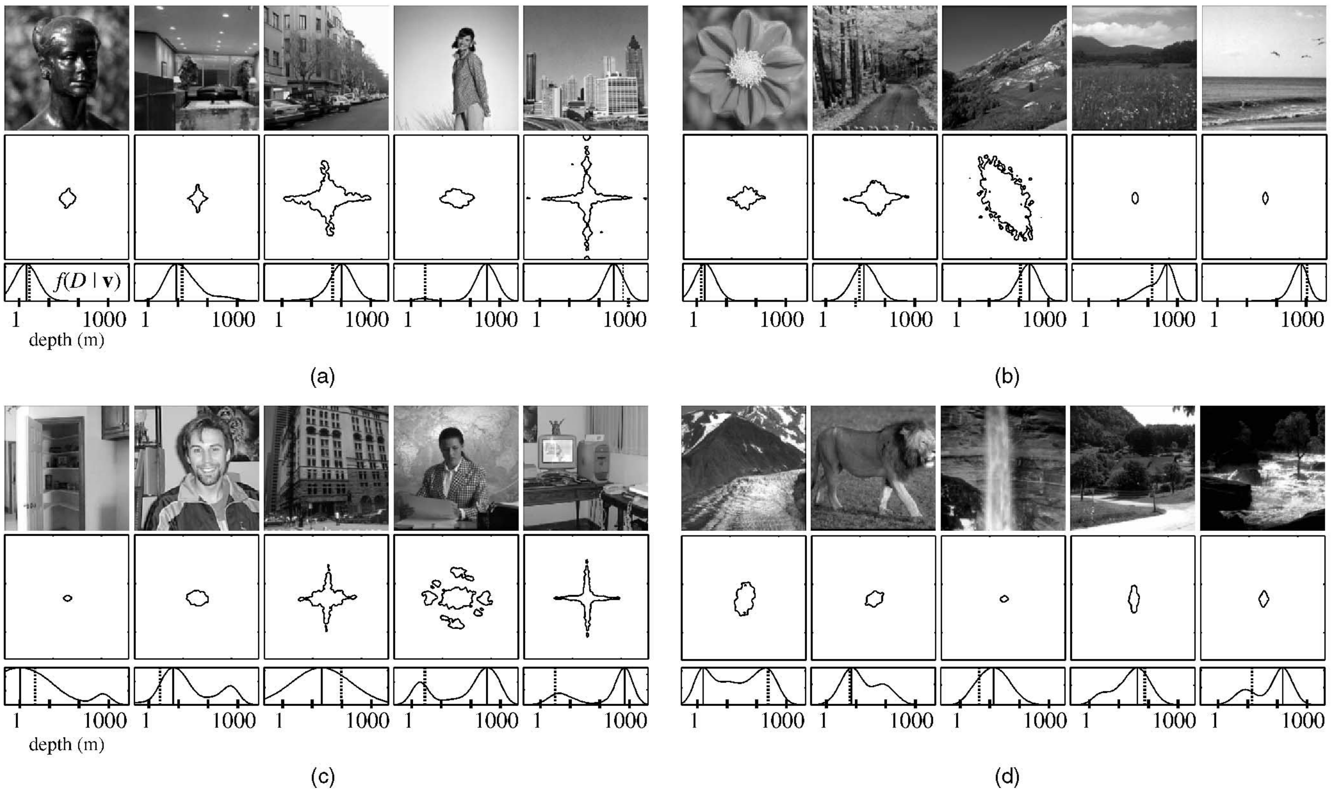
$$\mu_i^{k+1} = < \mathbf{v} >_i = \frac{\sum_{t=1}^{N_t} p^k(c_i \mid D_t, \mathbf{v}_t)\, \vec{v}_t}{\sum_{t=1}^{N_t} p^k(c_i \mid D_t, \mathbf{v}_t)}, \tag{17}$$

$$\mathbf{X}_i^{k+1} = < (\mathbf{v} - \mu_i^{k+1})(\mathbf{v} - \mu_i^{k+1})^T >_i, \tag{18}$$

$$\mathbf{b}_i^{k+1} = \left(\mathbf{X}_i^{k+1}\right)^{-1} < D\, \mathbf{v} >_i, \tag{19}$$

$$a_i^{k+1} = < D - \mathbf{v}^T \mathbf{b}_i^{k+1} >_i, \tag{20}$$

$$\sigma_i^{k+1} = < (D - a_i^{k+1} - \mathbf{v}^T \mathbf{b}_i^{k+1})^2 >_i\, . \tag{21}$$



(a) (b) (c) (d)

Fig. 7. Depth estimation using global energy features. For each picture we show the 80 percent section of its power spectrum, the function $f(D|\mathbf{v})$ obtained from the global energy features A_k, the estimated depth with the maximum likelihood (continuous line), and the real depth of the scene (dashed line). (a) and (b) Correspond to pictures selected among the 25 percent of pictures with the highest confidence and (c) and (d) pictures selected among the 25 percent of pictures with the lowest confidence. In (c) and (d), the PDFs have multimodal shape and produce more errors than in (a) and (b).

The simplified notation $<>_i$ represents the weighted average as detailed in (17). Further details concerning the learning equations may be found in [9]. As suggested in [9], we added a small constant to σ_i^{k+1} to the diagonal elements of $\mathbf{X}_i^{k+1}$ in order to avoid the clusters becoming too narrow. Restricting the matrices $\mathbf{X}_i$ to be diagonal slightly reduces performances but results in much fewer model parameters and faster learning. For the first iteration, the centers of the N_c clusters are initialized with the values of a random set of training data.

Once the learning is completed (after 30 iterations), the conditional PDF of depth, given the measured image features, is:

$$f_{D|v}(D\,|\,\mathbf{v},\,art) = \frac{\sum_{i=1}^{N_c} g(D\,|\,\mathbf{v}, c_i)\, g(\mathbf{v}\,|\,c_i)\, p(c_i)}{\sum_{i=1}^{N_c} g(\mathbf{v}\,|\,c_i)\, p(c_i)}. \tag{22}$$

Therefore, given a new scene picture, the mean depth is estimated from the image statistics as a mixture of linear regressions (11):

$$\hat{D} = \frac{\sum_{i=1}^{N_c} (a_i + \mathbf{v}^T \vec{b}_i)\, g(\mathbf{v}\,|\,c_i)\, p(c_i)}{\sum_{i=1}^{N_c} g(\mathbf{v}\,|\,c_i)\, p(c_i)}. \tag{23}$$

We can also estimate depth using the maximum likelihood: $\hat{D} = \max_D\{f_{D|v}(D\,|\,\mathbf{v},\,art)\}$. The estimation of the PDF $f(D|\mathbf{v}, S)$ provides a method to measure the reliability of the estimation provided by (23) for each new picture:

$$\sigma_D^2 = E[(\hat{D} - D)^2|\mathbf{v}] = \frac{\sum_{i=1}^{N_c} \sigma_i^2\, g(\mathbf{v}\,|\,c_i)\, p(c_i)}{\sum_{i=1}^{N_c} g(\mathbf{v}\,|\,c_i)\, p(c_i)}. \tag{24}$$

The confidence measure allows rejecting estimations that are not expected to be reliable. The bigger the value of the variance σ_D^2 the less reliable is the estimation $\hat{D}$.

6 Absolute Depth Estimation

In this section, we report the results of depth estimation using global and local structural features. The simulations were performed using a database of 8,000 images (256 × 256 pixels in size and 8 bits, gray scale). They come from the Corel stock photo library, a personal digital camera, images downloaded from the Web, and images captured from television. The database was composed of pictures of man-made and natural environments, close-up views of objects, and textures. Pictures with strange point of views were not included. The horizontal axis of the camera was parallel to the ground plane. In this study, color was not taken into account. Most of the images where in focus, therefore, blur cannot be used for depth measurements.

6.1 Depth Calibration

For the images used, the real distance and the aperture angle of the camera were unknown. Therefore, a calibration procedure was required in order to have absolute depth information for each picture. Depth was calibrated independently for man-made and natural scenes. The authors

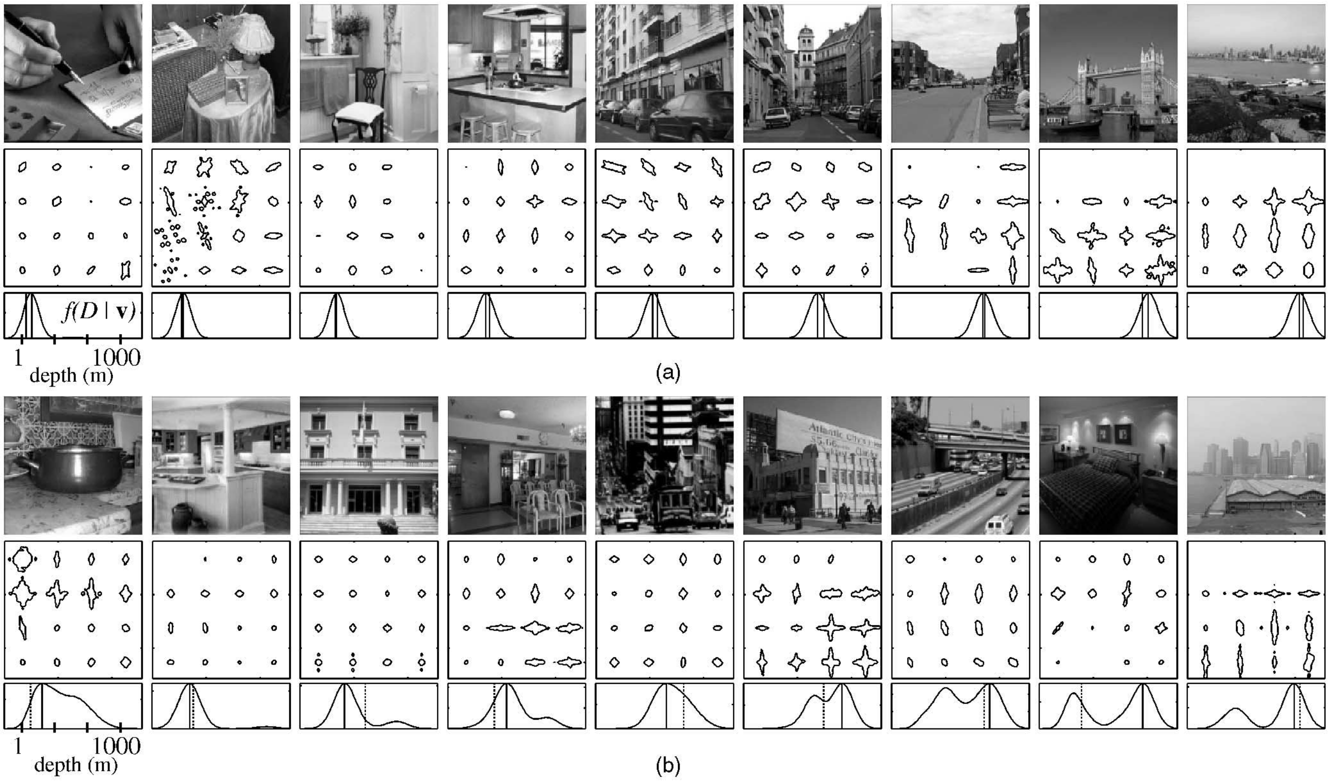


Fig. 8. Examples of man-made scenes sorted according to mean depth. (a) Pictures selected among the 85 percent of pictures with the highest confidence estimations. Middle line shows the 80 percent level section of the local amplitude spectrum and the bottom line shows the conditional PDF of depth obtained from the features $A_M(\mathbf{x}, k)$. (b) Pictures selected among the images with the lowest confidence estimations.

organized 4,000 images of each group according to their mean depth. Then, four observers reported the depth, in meters, of 400 images. A third order polynomial was used to fit the reported depths, averaged across the four subjects, as function of the ranks obtained in the sorting task.

6.2 Depth from Global Image Statistics

For the learning of the relationship between depth and global structural features, $f(D, \mathbf{v} \mid S)$, we used 2,000 images for the training for each group, $S = \{art, nat\}$ ranging from close-up to panoramic views. We tested the ability to predict the scene scale of new pictures using the global energy features (8), A_k and the magnitude correlation features (9), $A_{i,j}$. In both cases, the features vector was given by projecting the representations into their respective principal components (Section 4). In both cases, the best performance were achieved when using $N_c = 8$ clusters and about $L = 25$ PCs. The images were decomposed using a bank of Gabor filters tuned to six orientations and five scales ($K = 30$). Increasing the complexity of the model did not result in an increase of performance.

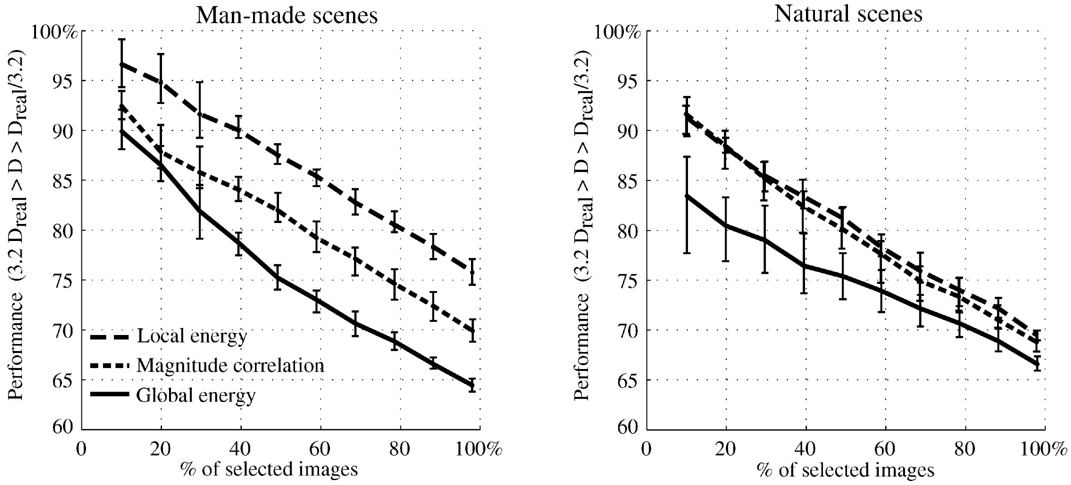
Fig. 7 shows pictures of man-made and natural scenes with their amplitude spectrum and the estimated conditional PDFs (22) when using the PCs of the global energy features, A_k. From the PDFs, for each image we estimated the mean depth of the scene (23) and the confidence of the estimation (24). Fig. 9 compares the performance of the two sets of features and shows how performance increases when rejecting images with low confidence estimations. For man-made scenes, when considering all the images from the test set, the estimated depth was in the same decade ($\hat{D}_{art} \in [D_{real}/3.2, D_{real} * 3.2]$) than the real depth for 65 percent of the pictures when using the global energy features. Performances were better when using the PCs of the magnitude correlation features (70 percent). For these two sets of measures, the number of features used for the estimation is the same ($L = 25$) as we used the same number of PCs for training the algorithm. Similar results were obtained for natural scenes.

Performance increase when rejecting images with low confidence. Figs. 7a and 7b show a set of images selected among the 25 percent of the images from the test set that provided the highest confidence levels using the global energy features. Figs. 7c and 7d show examples of man-made and natural scenes with low confidence estimations.

Although a mean depth estimation based on global features is unreliable for most of the pictures, there is a significant correlation between unlocalized structure and absolute depth (0.64 for man-made scenes and 0.72 for natural scenes) showing that simple image statistics vary with the real scene scale. To obtain reliable estimations, we have to include spatial information.

6.3 Depth from Local Features

Fig. 8 shows estimation results obtained using the local energy of the wavelet coefficients (10), $A_M(\mathbf{x}, k)$. The spatial resolution of $A_M(\mathbf{x}, k)$ was set to $M = 4$ as it provides the best performance. Increasing the resolution did not improve the results. For the learning, we used 2,000 images, $N_c = 8$ clusters and $L = 25$ PCs. Fig. 9 compares performance using the local energy features to the estimation using features from global image statistics. Performances are significantly better for man-made scenes but not for natural scenes. For man-made



(a) (b)

Fig. 9. Results of depth estimation for man-made and natural scenes. The graphs compare the performances of the three sets of features (global energy, A_k; magnitude correlation, $A_{i,j}$; and local energy; $A_M(\mathbf{x}, k)$ with $M = 4$) and show how performances increase when considering images with higher confidence (smaller σ_D). Performance indicates the percentage of images with estimated depth in the interval $\hat{D}_{art} \in [D_{real}/3.2, D_{real} * 3.2]$. The error bars represent 90 percent of variability when using different sets of images for the training and for the test. In all the graphs we used, $N_c = 8$ clusters to model the PDFs and $L = 25$ features extracted from the PCA for each of the three set of measures.

scenes, 76 percent of depth estimations were in the interval $\hat{D}_{art} \in [D_{real}/3.2, D_{real} * 3.2]$. Performances rapidly increase when considering images with high confidence estimations. Performances reach 88 percent when considering, for man-made scenes, 50 percent of the pictures from the test set with the highest confidence (Figs. 8a and 10). For natural images the results were lower. 70 percent of the estimations $\hat{D}_{nat}$ were in the interval $\hat{D}_{nat} \in [D_{real}/3.2, D_{real} * 3.2]$ which is the same result than the one obtained using the magnitude correlation features.

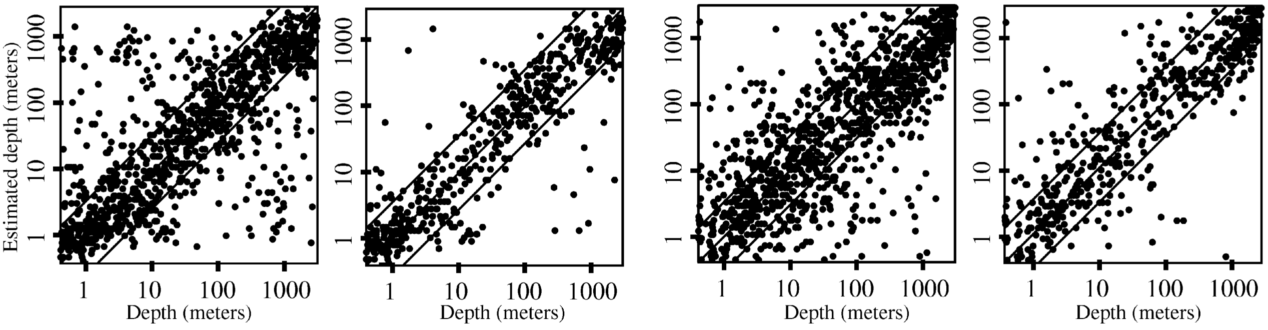
In order to compare the algorithm performance with human performance, we compared the depth reported by each observer (as described in Section 5.1) with the mean depth reported by the others as a measure of consistency between different observers. For man-made scenes, the percent of pictures with an error in the interval of one decade with respect to the mean, for each subject, was 82 percent, 90 percent, 93 percent, 81 percent, and 79 percent (the first subject is one author and the others are nontrained subjects). Results are slightly better than those provided by the algorithm (76 percent). For natural scenes, the percent of pictures with an error in the interval of one decade was 74 percent, 79 percent, 61 percent, 66 percent, and 54 percent which is similar to the performance obtained by the local energy statistics (Fig. 9).

It is interesting to consider independently the structures present at different locations in the image by writing the conditional PDF as:

$$f(A_M(\mathbf{x}, k) \,|\, D, \mathit{art}) = \prod_{j=1}^{N} f_{x_j,y_j}(\mathbf{v}_{x_j,y_j} \,|\, D, \mathit{art}) \qquad (25)$$

which corresponds to consider that, once D is specified, the image features at different spatial locations, are independent. At each location $\mathbf{x}_j = (x_j, y_j)$, the vector of local features is $\mathbf{v}_{\mathbf{x}_j} = \{A_M(\mathbf{x}_j, k)\}_{k=1,K}$. Therefore, $\mathbf{v}_{\mathbf{x}_j}$ contains the energy of the wavelet coefficients averaged over a neighborhood of the location $\mathbf{x}_j$.

The conditional PDFs $f_{x_j,y_j}(\mathbf{v}_{x_j,y_j} \,|\, D, \mathit{art})$ model the statistical dependencies between the energy features $\mathbf{v}_{x_j,y_j}$ at the spatial location x_j, y_j and the mean depth of the scene D. Each local conditional PDF can be estimated independently



(a) (b) (c) (d)

Fig. 10. Estimation results on 1,000 images from the test set using local second order image statistics for man-made (a) and (b) and natural scenes (c) and (d). (b) and (d) show the results for the 50 percent of the images of each set with the highest confidence level.

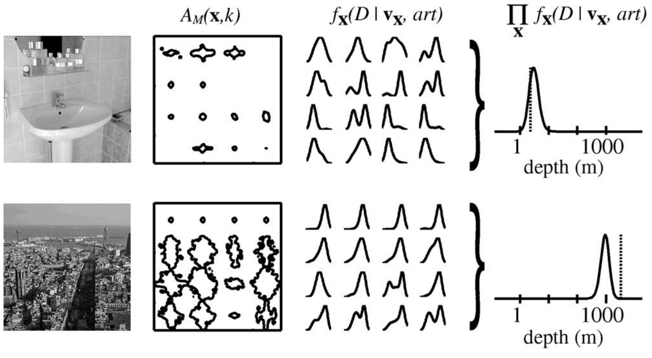

$A_M(\mathbf{x},k)$
$f_\mathbf{x}(D \mid \mathbf{v_x}, art)$
$\prod_\mathbf{x} f_\mathbf{x}(D \mid \mathbf{v_x}, art)$
1
1000
depth (m)
1
1000
depth (m)

Fig. 11. Examples of scenes with man-made structures and the relation between local structures ($M = 4$) and the mean depth of the scene. Note that similar structures are associated to different depths as a function of image position.

with the model of (12) and trained using the EM algorithm. We assume a uniform prior for the distributions of depths $f(D)$ in the database. For the training, we used $N_c = 8$ clusters to model each local PDF. Due to the nonstationary statistics of real-world images, the functions f_{x_j,y_j} will differ from one spatial location to another. Fig. 11 shows two examples of depth estimation using (25). Performance were the same as the one reported in Fig. 9 using the PCs of $A_M(\mathbf{x}, k)$. Although this procedure results in more computations than when using the features obtained by the PCA, the writing of the PDF as in (25) allows studying how the local structures relate to the mean depth of the scene.

There are some important observations about the relationship between the local image structure and the mean depth of the scene:

- For the features used, the relationship between the mean depth of the scene and the local structure at one spatial position is weak and it is necessary to integrate information across the whole picture to improve the reliability. Fig. 11 shows the shape of the local conditional PDFs f_{x_j,y_j} of depth given the local structure present in the picture. Most of the local PDFs are multimodal or spread across several orders of magnitude in depth.
- The functions f_{x_j,y_j} model the relationship between the structure at one spatial location and the mean depth of the whole scene. They are not necessarily related to the local depth at the location around (x_j, y_j). For example, the texture associated with the sky indicates mean depths in the range from 100 meters to panoramic, which does not correspond to the distance between the observer and the clouds.
- The same local structure $(\mathbf{v_x})$ can be related to different mean depths when located at different spatial positions (Fig. 11). A flat surface in the top will be interpreted as a possible sky patch. However the same surface in the bottom part will be an indicator of short and mean distances (from 1 to 100 meters). On the contrary, a textured patch located in the top part will discard far depths, but located in the bottom, it may be correlated with a panoramic view (see Fig. 11).
- As suggested by Fig. 5, the image statistics are stationary with respect to the horizontal spatial dimension, x. That means that the local conditional PDFs can be written as $f_{x_j,y_j} = f_{y_j}$. The results remain unaffected by this simplification. However, scene structure statistics and their relationship with the mean depth of the scene strongly vary with respect to elevation, y_j, and consequently this is an important factor for depth estimation. Assuming complete stationary, $f_{x_j,y_j} = f$, gives poor estimation results.

To summarize, the results show that reliable estimations of the absolute depth of a real-world scene may be computed from monocular information by recognizing structures and texture patterns in the image. In the next section we introduce some applications of computing the mean depth of a scene.

7 Application for Scene Recognition and Object Detection

Computing the complete 3D structure of the scene yields a great amount of information useful for motion planning, grasping objects, etc. However, a coarse estimate of the mean distance between the observer and the background and main objects composing the scene is relevant for identifying the context in which the observer is immersed and can be used to restrict the search and recognition of objects.

7.1 Scene Category Recognition

Scene category recognition refers to the classification of a scene into a semantic group (e.g. street, room, forest, etc.). With the development of applications in image indexing, novel procedures in computational scene recognition have been recently proposed ([3], [5], [7], [10], [22], [36], [39]), but recognition performances are limited by the small number of semantic categories that these models propose (e.g., city versus landscape, indoor versus outdoor, suburban versus urban scenes). In that regard, adding the estimation of the mean depth of a scene to other attributes may significantly increase performances of semantic recognition. As an illustration, Fig. 12 shows the distribution, along the mean depth axis, of basic scene categories commonly employed by human observers when asked to name images [22], [29], [39]. Even if the groups overlap, the mean depth allows the emergence of specific semantic categories, like objects, indoors, urban streets, highways and panoramic environments for man-made structures, and rivers/forests, fields, mountains, and ocean views for natural images.

7.2 Task and Context-Dependent Scale Selection

One fundamental problem in computational vision is to find which are the scales in which the main elements of the scene are localized in the picture. If this information is available as a result of a low cost preprocessing stage, then subsequent stages of object detection and recognition could be greatly simplified by focusing the processing onto the only diagnostic/relevant scales. In that aim, Lindeberg [18], [19] proposed a method for scale selection for the detection of low-level features as edges, junctions, ridges, and blobs when there is no a priori information about the nature of the picture. The method is based on the study of the evolution over scales of scale-space derivatives.

We propose to use the mean depth to select the scale at which one particular object can be found [38]. This provides

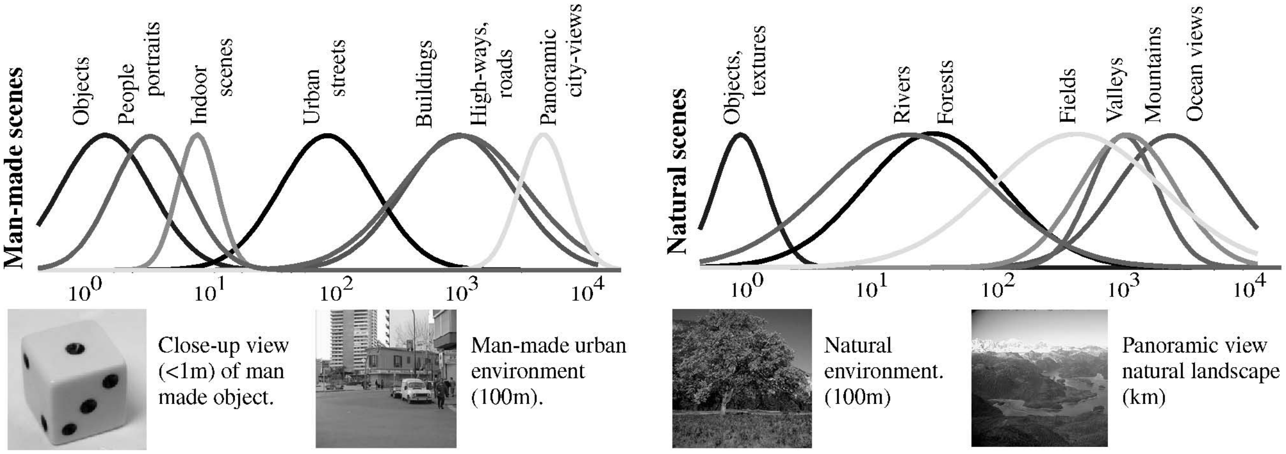


Fig. 12. Distribution of scene categories commonly used by subjects in function of the mean depth of the scene for man-made and natural structures. These distributions are obtained from a separate annotation (see [22]). The distributions are then obtained by fitting a Gaussian distribution to the PDF $f(D|category)$ for each scene category. The bottom examples show images and their descriptions generated using only the mean depth and the discrimination between natural and man-made scenes.

a method for scale selection that is both task and context dependent. The expected size of one object can be estimated by using the mean depth of the scene by $\hat{S}_{obj} \simeq K_{obj}/\hat{D}^2$, were K_{obj} is a normalization constant. This will give a restriction of the possible scales as $\hat{D}$ here refers to the estimation of the mean depth of the scene. Fig. 13 shows the expected sizes of heads for different scenes computed using the mean depth given by the algorithm.

We selected a subset of pictures in man-made environments containing people (urban outdoor and indoor environments from 1 to 100 meters). We trained the algorithm to predict the height of the people's heads based in the local structural information. For 83 percent of the scenes tested (900 for the learning and 250 for the test), the estimated height of people's heads was in the interval $[H/2, H * 2]$, where H is the true height measured directly from the image. As a consequence, the estimated distance ($\hat{D} = K/\hat{H}$) is also in the interval $[D/2, D * 2]$ for 83 percent of scenes tested. This result is better than the one reported in Section 6.3 due to be working with only a subset of scenes (scenes with people) and to have a more accurate depth calibration based on head height and not on subjective evaluations of mean depth. Note that the estimated distance ($\hat{D}$) of people in the pictures is obtained without any detection stage of faces or bodies. Instead, we use the whole picture as the context in which heads are located [38].

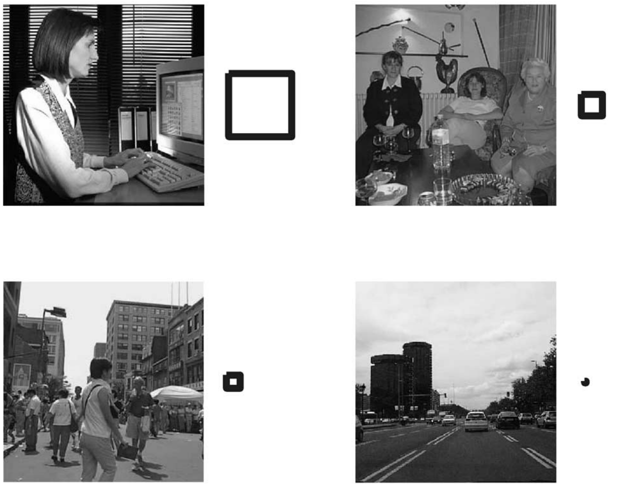


Fig. 13. Task and context-dependent scale selection. The boxes represent the attended size of people heads estimated using the mean depth as a descriptor of context information.

8 Conclusion

The dominant structures present in a scene (e.g., squared blocks, diagonal planes, horizontal surfaces, small grain textures, etc.) that confer to a space its identity (e.g., a highway, a street, a coast) strongly differ with spatial scale. In other words, the structure of the space, the size and position of the main elements of the scene vary with the distance of the observer (spatial scale) in a very predictable and regular way. The results of this paper show that:

- There exist differential structural regularities at different scales in both man-made and natural environments. Therefore, natural and man-made real-world structures are not self-similar when we change the scale of analysis.
- Those structural regularities are stable enough to estimate the absolute mean depth of a scene by recognizing the structures present in the image.

Depth computation as proposed here does not require recovering the local 3D structure of the scene as an intermediate step. The recognition of structures in the scene provides absolute depth related information that does not require object recognition, processing of surfaces, shading, or junctions. Therefore, the estimated depth provides contextual information and can be used to simplify object recognition stages by choosing the more adequate scale of analysis and by limiting the type of possible objects. Furthermore, mean depth is a key attribute for scene recognition. Combined with other perceptual attributes [22], depth can allow the recognition of the semantic category of the scene as a first step in the visual processing before the analysis of 3D surfaces or object detection.

ACKNOWLEDGMENTS

The authors would especially like to thank Professor W. Richards for fruitful discussions and three anonymous reviewers whose comments helped improve the manuscript. Thanks also to Javid Sadr for useful comments about the manuscript.

REFERENCES

[1] R. Baddeley, "The Correlational Structure of Natural Images and the Calibration of Spatial Representations," *Cognitive Science,* vol. 21, pp. 351-372, 1997.

[2] H.G. Barrow and J.M. Tenenbaum, "Interpreting Line Drawings as Tree-Dimensional Surfaces," *Artificial Intelligence,* vol. 17, pp. 75-116, 1981.

[3] K. Barnard and D.A. Forsyth, "Learning the Semantics of Words and Pictures," *Proc. Int'l Conf. Computer Vision,* vol. 2, pp. 408-415, 2001.

[4] J.R. Bergen and M.S. Landy, "Computational Modeling of Visual Texture Segregation," *Computational Models of Visual Processing,* M.S. Landy and J.A. Movshon, eds., pp. 253-271, Cambridge, Mass.: MIT Press, 1991.

[5] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Region-Based Image Querying," *Proc. IEEE Workshop Content-Based Access of Image and Video Libraries,* pp. 42-49, 1997.

[6] J.M. Coughlan and A.L. Yuille, "Manhattan World: Compass Direction from a Single Image by Bayesian Inference," *Proc. Int'l Conf. Computer Vision,* pp. 941-947, 1999.

[7] J.S. De Bonet and P. Viola, "Structure Driven Image Database Retrieval," *Advances in Neural Information Processing,* vol. 10, pp. 866-872, 1997.

[8] D.J. Field, "Relations between the Statistics of Natural Images and the Response Properties of Cortical Cells," *J. Optical Soc. Am.,* vol. 4, pp. 2379-2394, 1987.

[9] N. Gershnfeld, *The Nature of Mathematical Modeling,* Cambridge Univ. Press, 1999.

[10] M.M. Gorkani and R.W. Picard, "Texture Orientation for Sorting Photos at a Glance," *Proc. Int'l Conf. Pattern Recognition,* vol. 1, pp. 459-464, 1994.

[11] P.J. Hancock, R.J. Baddeley, and L.S. Smith, "The Principal Components of Natural Images," *Network,* vol. 3, pp. 61-70, 1992.

[12] D.J. Heeger and J.R. Bergen, "Pyramid Based Texture Analysis/Synthesis," *Proc. ACM SIGGRAPH, Computer Graphics,* pp. 229-238, 1995.

[13] B.K.P. Horn and M.J. Brooks, *Shape from Shading,* Cambridge, Mass.: MIT Press, 1989.

[14] A. Jepson, W. Richards, and D. Knill, "Modal Structures and Reliable Inference," *Perception as Bayesian Inference,* D. Knill and W. Richards, eds., pp. 63-92, Cambridge Univ. Press, 1996.

[15] M.I. Jordan and R.A. Jacobs, "Hierarchical Mixtures of Experts and the EM Algorithm," *Neural Computation,* vol. 6, pp. 181-214, 1994.

[16] J.M. Keller, R.M. Crownover, and R.Y. Chen, "Characteristics of Natural Scenes Related to the Fractal Dimension," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 9, no. 5, pp. 621-627, 1987.

[17] A.B. Lee, D. Mumford, and J. Huang, "Occlusion Models for Natural Images: A Statistical Study of a Scale-Invariant Dead Leaves Model," *Int'l J. Computer Vision,* vol. 41, no. 1/2, pp. 35-59, 2001.

[18] T. Lindeberg, "Detecting Salient Blob-Like Image Structures and Their Scales with a Scale-Space Primal Sketch: A Method for Focus-of-Attention," *Int'l J. Computer Vision,* vol. 11, no. 3, pp. 283-318, 1993.

[19] T. Lindeberg, "Principles for Automatic Scale Selection," *Int'l J. Computer Vision,* vol. 30, no. 2, pp. 77-116, 1998.

[20] F. Liu and R.W. Picard, "Periodicity, Directionality, and Randomness: Wold Features for Image Modeling and Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, pp. 722-733, 1996.

[21] A. Oliva, A. Torralba, A. Guerin-Dugue, and J. Herault, "Global Semantic Classification Using Power Spectrum Templates," *Proc. Challenge of Image Retrieval,* 1999.

[22] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," *Int'l J. Computer Vision,* vol. 42, no. 3, pp. 145-175, 2001.

[23] B.A. Olshausen and D.J. Field, "Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images," *Nature,* vol. 381, pp. 607-609, 1996.

[24] S.E. Palmer, *Vision Science,* Cambridge, Mass.: MIT Press, 1999.

[25] A. Papoulis, *Probability, Random Variables and Stochastic Processes,* second ed. MacGraw-Hill, 1984.

[26] A.P. Pentland, "Fractal-Based Description of Natural Scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, pp. 661-674, 1984.

[27] J. Portilla and E.P. Simoncelli, "A Parametric Texture Model Based on Joint Statistics of Complex Wavelets Coefficients," *Int'l J. Computer Vision,* vol. 40, pp. 49-71, 2000.

[28] D.L. Ruderman, "Origins of Scaling in Natural Images," *Vision Research,* vol. 37, pp. 3385-3398, 1997.

[29] B.E. Rogowitz, T. Frese, J.R. Smith, C. Bouman, and E. Kalin, "Perceptual Image Similarity Experiments," *Proc. SPIE, Conf. Human Vision and Electronic Imaging,* Jan. 1998.

[30] A. van der Schaaf and J.H. van Hateren, "Modeling of the Power Spectra of Natural Images: Statistics and Information," *Vision Research,* vol. 36, pp. 2759-2770, 1996.

[31] B. Schiele and J.L. Crowley, "Recognition without Correspondence Using Multidimensional Receptive Field Histograms," *Int'l J. Computer Vision,* vol. 36, no. 1, pp. 31-50, 2000.

[32] I. Shimshoni, Y. Moses, and M. Lindenbaum, "Shape Reconstruction of 3D Bilaterally Symmetric Surfaces," *Int'l J. Computer Vision,* vol. 2, pp. 1-15, 2000.

[33] E.P. Simoncelli and W.T. Freeman, "The Steerable Pyramid: A Flexible Architecture for Multi-Scale Derivative Computation," *Proc. Second IEEE Int'l Conf. Image Processing,* Oct. 1995.

[34] E.P. Simoncelli and B.A. Olshausen, "Natural Image Statistics and Neural Representation," *Ann. Rev. Neuroscience,* vol. 24, pp. 1193-1216, 2001.

[35] B.J. Super and A.C. Bovik, "Shape from Texture Using Local Spectral Moments," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 4, pp. 333-343, 1995.

[36] M. Szummer and R.W. Picard, "Indoor-Outdoor Image Classification," *Proc. IEEE Int'l Workshop Content-Based Access of Image and Video Databases,* 1998.

[37] A. Torralba and A. Oliva, "Scene Organization Using Discriminant Structural Templates," *Proc. Int'l Conf. Computer Vision,* pp. 1253-1258, 1999.

[38] A. Torralba and P. Sinha, "Statistical Context Priming for Object Detection: Scale Selection and Focus of Attention," *Proc. Int'l Conf. Computer Vision,* vol. 1, pp. 763-770, 2001.

[39] A. Vailaya, A. Jain, and H.J. Zhang, "On Image Classification: City Images vs. Landscapes," *Pattern Recognition,* vol. 31, pp. 1921-1935, 1998.

[40] S.C. Zhu, Y. Wu, and D. Mumford, "Filters, Random Fields and Maximum Entropy (FRAME)," *Int'l J. Computer Vision,* vol. 27, no. 2, pp. 1-20, 1998.

Antonio Torralba received the telecommunication engineer degree from the Universitat Politecnica de Catalunya, Barcelona, Spain, in 1995 and the MS degree (1996) and the PhD degree (1999) in signal and image processing from the Institut National Polytechnique de Grenoble, France. In 2000, he began postdoctoral work in the Department of Brain and Cognitive Sciences at MIT and he is currently at the MIT Artificial Intelligence Laboratory. His research interests include natural image statistics, scene recognition, and contextual models of object recognition.

Aude Oliva received the MS degree (1991) in psychology from University P.M. France, Grenoble and the MS (1992) and PhD (1995) degrees in cognitive science, from the Institut National Polytechnique de Grenoble (INPG), France. In 2000, after research positions in the UK, Japan and France, she joined the Center for Ophthalmic Research, at Harvard Medical School, Brigham and Women's Hospital, Boston. Her research focuses on real world image perception, recognition and memory, as well as the dynamic aspects of attention and visual search in scene understanding.