massachusetts institute of technology — artificial intelligence laboratory

Global Depth Perception from Familiar Scene Structure

Antonio Torralba and Aude Oliva

AI Memo 2001-036 December 2001
CBCL Memo 213

© 2001 massachusetts institute of technology, cambridge, ma 02139 usa — www.ai.mit.edu

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 inform about the actual 'size' of the space. One possible source of information for absolute depth estimation is the image size of known objects. However, this is computationally complex due to the difficulty of the object recognition process. Here we propose a source of information for absolute depth estimation that does not rely on specific objects: we introduce a procedure for absolute depth estimation based on the recognition of the whole scene. The shape of the space of the scene and the structures present in the scene are strongly related to the scale of observation. 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.

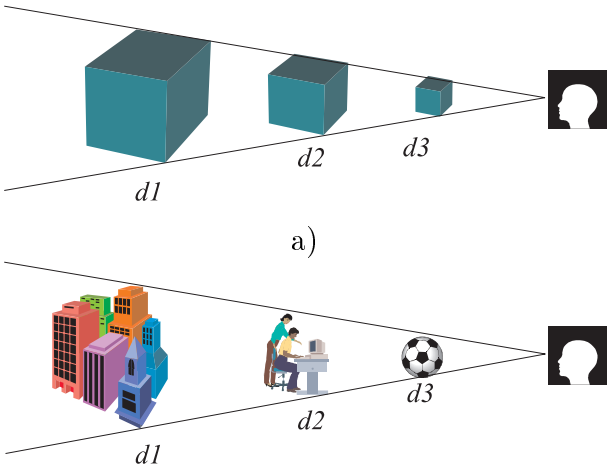

Fig. 1. a) Artificial stimulus: the monocular information cannot provide an absolute depth percept. b) Real-world stimulus: the recognition of familiar image structures provides unambiguous monocular information about the absolute depth between the observer and the scene.

I. Introduction

Figure 1.a. illustrates the fundamental problem of depth perception from monocular information. In the absence of cues for absolute depth measurement 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 however apply when dealing with real-world stimuli (figure 1.b). Physical processes that shape natural structures are different at each scale (e.g., leaves, forests, mountains). Human kind also builds 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 [11].

The constraints on the structure of the 3D scene at each spatial scale can be directly transposed into image content. Figure 2 shows three pictures representing environments with different depths: the scenes strongly differ in their global configuration, the size of the component surfaces as well as 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 defining smaller surfaces. Close-up views of objects tend to have large flat surfaces and, on average, no clear dominant orientations. 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 [10], from texture gradients [27], from edges and junctions [2], from symmetric patterns [25], and from other pictorial cues such as occlusions, relative size, and elevation with respect the horizon line (see [19] for a more complete review). 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: binocular vision, motion parallax and defocus. Absolute depth measurements based on these techniques can only be achieved when particular sources of information are available (multiple cameras, motion and defocus). However, under normal viewing conditions, human subjects have no problem in providing a rough estimate of the absolute depth of the scene even in the absence of all these sources of information (for instance, when looking at a picture of

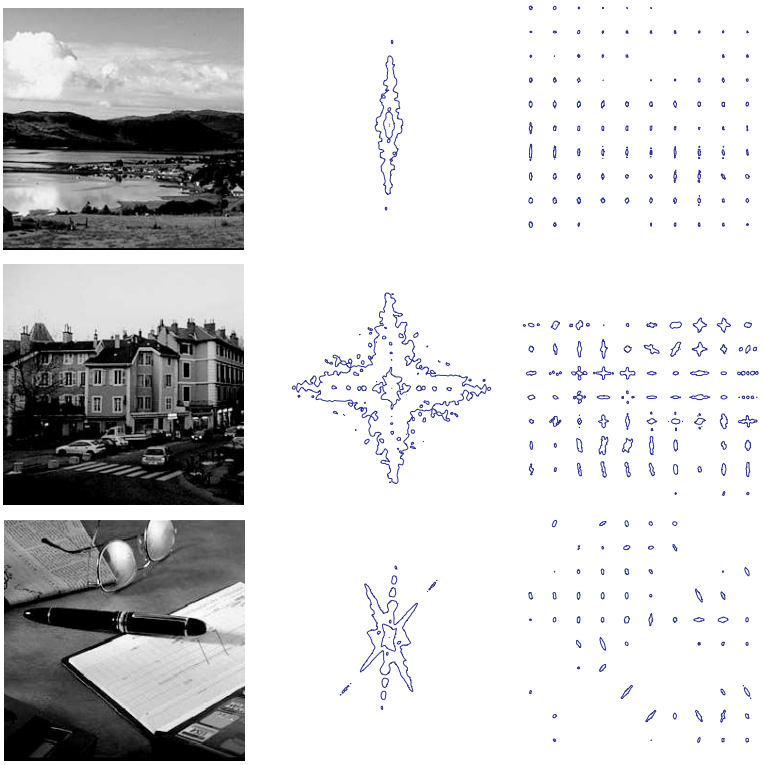

Fig. 2. Three examples of images used in this study. The scenes strongly differ in their absolute mean depth. For each scene, we show the sections of the global magnitude of the Fourier transform (center) and the sections of the magnitude of the windowed Fourier transform (right).

a real-world scene). One additional source of information for absolute depth estimation is the image size of familiar objects with fixed size as faces, bodies, cars, etc. However, this requires exploring the image into detail and decomposing it into its constituent elements (objects, textures, surfaces, etc.). 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 new source of information for absolute depth computation from monocular views: the familiar global image structure. We propose a procedure for absolute depth estimation based on the recognition of the structure of the scene image. The underlying hypothesis of this approach is that the recognition of the scene as a whole (categorization of the scene as an indoor, a street, a panoramic view of a landscape, etc.) is a simpler problem than the one of general object detection and recognition [17], [18], [29].

II. Scene Structure

In the last years, there have been an increasing number of models of the image structure based on textural features mainly motivated by applications in image indexing and computation of similarities between pictures that match human criteria ([4], [5], [8], [16], [31] among many others). For the purpose of this section, we consider a simple definition of the image structure based on a Fourier description of the textural patterns present in the image and their spatial arrangement [17], [18], [29]. The Fourier transform is one of the basic tools for the description of image textural patterns. It provides information about the pattern of dominant orientations and the global roughness of the texture (e.g. [3]). In this section, we present two levels of description of the image structure (figure 2) based on the Fourier transform. 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 the local windowed Fourier transform, provides the

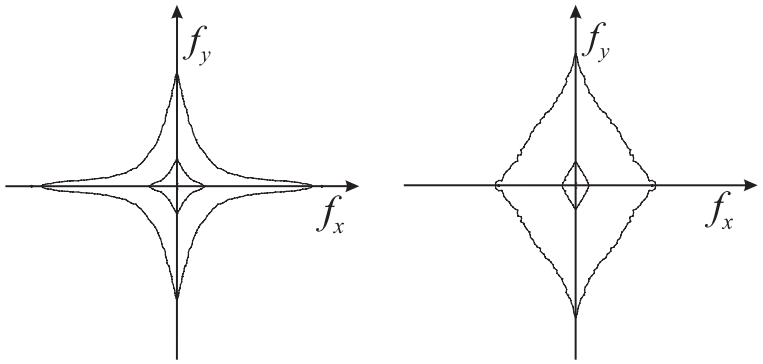

Fig. 3. Global spectral signatures of man-made and natural environments computed on 6000 images. The contour plots represent the 50% and the 80% of the energy of the spectral signatures.

dominant orientations and scales with a coarse description of their spatial distribution. The aim of this section is to study the regularities of these two structural descriptions in order to inform about the mean depth of a real world scene.

A. Unlocalized scene structure and spectral signatures

The discrete Fourier transform of an image is defined as:

$$I(f_x, f_y) = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} i(x,y)\, h(x,y)\, e^{-j\,2\,\pi(f_x x + f_y y)} \tag{1}$$

where $i(x,y)$ is the intensity distribution of the image along the spatial variables (x,y); $j = \sqrt{-1}$; f_x and f_y are the spatial frequency variables defined in $(f_x, f_y) \in [-0.5, 0.5] \times [-0.5, 0.5]$ (units are in cycles per pixel); and $h(x,y)$ is a circular symmetric window that reduces the boundary effects. For the results shown here, the pictures are square images with 256x256 pixels. The amplitude spectrum is defined as the magnitude of the Fourier transform: $A(f_x, f_y) = |I(f_x, f_y)|$. Figure 2 shows examples of sections of the amplitude spectra of scene pictures. The amplitude spectrum reveals the dominant orientations and textural patterns in the image. 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 Fourier transform. Therefore, the amplitude spectrum contains no information about the local arrangement of edges, surfaces, and objects inside the image. In fact, if we consider an image as being any possible distribution of pixel intensities, then the amplitude spectrum is not informative because many images would have the same amplitude spectrum. However, in the context of real-world scene pictures, the shape of the amplitude spectrum is correlated with the shape of the spatial structure of the scene picture [18]. In order to study the relationship between the amplitude spectrum and the scene structure, we define the *spectral signature* of a set of images S as the mean amplitude spectrum:

$$\overline{A}_S(f_x, f_y) = E\,[\,A(f_x, f_y)\,|\,S] \tag{2}$$

where E is the expectation operator. The spectral signature of a set of images reveals the dominant structures shared by the images of the set S. Several studies [7], [23] have observed that the averaged amplitude spectrum of the set of real-world scene pictures falls with a form: $\overline{A}_S \sim 1/f^{\alpha}$ with $\alpha \sim 1$. Real-world scenes can be divided into semantic categories that depict specific spectral signatures (see [17] for a detailed discussion). The clearest example of picture sets opposed by their spectral signatures is man-made vs. natural structures. Both spectral signatures are defined by the conditional expectations:

$$\overline{A}_{art}(f_x, f_y) = E\,[\,A(f_x, f_y)\,|\,man - made\,] \tag{3}$$

$$\overline{A}_{nat}(f_x, f_y) = E\,[\,A(f_x, f_y)\,|\,natural\,] \tag{4}$$

Figure 3 shows the contour plots of the spectral signatures obtained from more than 6000 pictures[1]. $\overline{A}_{art}(f_x, f_y)$ has dominant horizontal and vertical orientations due to the bias found in man-made structures [1], [18]. $\overline{A}_{nat}(f_x, f_y)$ 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 landscapes with a very high confidence (93.5%, refer to [18], [29]).

B. Windowed Fourier transform and local spectral signatures

An essential aspect of a scene description, that the global amplitude spectrum does not encode, concerns the spatial arrangement of the structural elements (salient edges and textures). For example, panoramic landscapes have the sky at the top, characterized by low spatial frequencies, the horizon line (revealing vertical frequency component), and usually texture at the bottom part. Typical 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 (e.g. [24]). That specific arrangement produces a particular spatial pattern of dominant orientations and scales (figure 2) that can be described using the magnitude of the Windowed Fourier transform (WFT). The WFT is defined as:

$$I(x, y, f_x, f_y) = \sum_{x',y'=0}^{N-1} i(x', y')\, h_r(x' - x, y' - y)\, e^{-j\,2\pi(f_x x' + f_y y')} \tag{5}$$

$h_r(x', y')$ is a Hamming window with a circular support of radius r. The local amplitude spectrum at one spatial location (x, y) is: $A(x, y, f_x, f_y) = |I(x, y, f_x, f_y)|$. The function $A(x, y, f_x, f_y)$ encodes the local image structure in a neighborhood of the location (x, y) (figure 2). Due to the size of the window h_r, the local amplitude spectrum only contains the low spatial frequencies of the distribution of the structures in the scene. The *local spectral signature* of a set of images S is defined as follows:

$$\overline{A}_{S,r}(x, y, f_x, f_y) = E\,[\,A(x, y, f_x, f_y)\,|\,S,\,r\,] \tag{6}$$

The local spectral signature of a set of images gives information about the dominant spectral features and their mean spatial distribution. Figure 4 shows the local spectral signatures ($r = 16$ pixels) for pictures of man-made and natural environments. As natural and man-made environmental categories cover (almost) all the possible spatial arrangements of their main components, the local spectral signatures are spatially homogeneous (the second order statistics are stationary). However, this is not the case when defining other scene categories (see section III and [18]).

III. Scene structure as a depth cue

The aim of this section is to justify the use of the spectral features for describing the scene structure and to present the main sources of variability in the spectral features with respect to depth. Our main argument is that there are strong similarities in the spectral components (global and local) of scenes sharing the same mean depth (spatial scale). There are at least two reasons that can explain the dependence between the scene structure and its mean depth:

[1]Images were all 256 x 256 pixels in size, in 256 gray levels. They come from the Corel stock photo library, pictures taken from a digital camera, images downloaded from the web and images captured from TV. The scene database was composed of scenes of natural environments (e.g. coast, beach, ocean, island, field, desert, grassland, valley, lake, river, mountains, canyon, cavern, forest, waterfall, garden, etc.), man-made environments (e. g. skyscraper, city center, commercial area, street, road, highway, house, building, pedestrian center, place, parking, indoors etc.) and close-up views of objects and textures. Pictures with strange point of views were not included. The horizontal axis of the camera is, more or less, parallel to the ground plane. If we allow random rotations of the camera, then the spectral signatures will lose their characteristic orientation pattern that is due to the bias of edges aligned with respect to the gravity plane. Also pictures are taken at a height corresponding to human size (there are no pictures looking from above or satellite views).

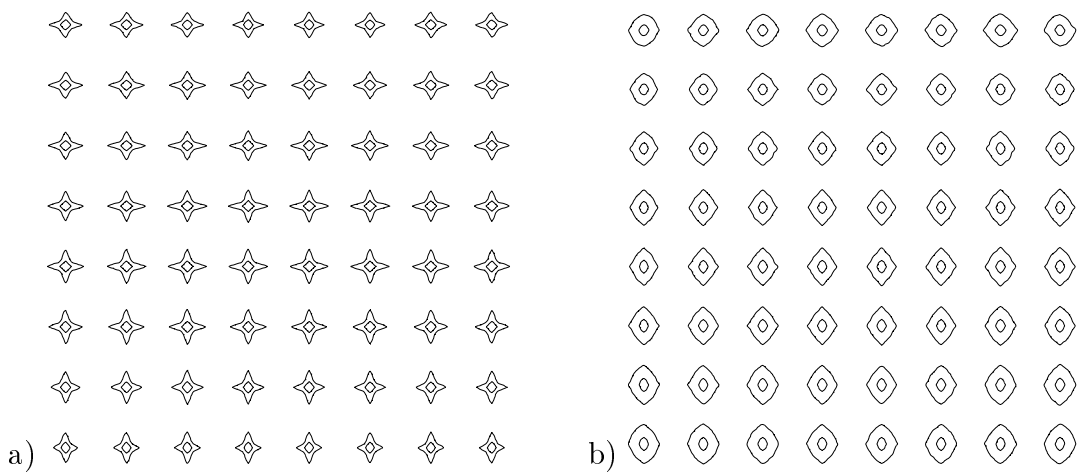

Fig. 4. Local spectral signatures of (a) man-made and (b) natural environments computed on 6000 images. Here, the function $\overline{A}_{S,r}(x, y, f_x, f_y)$ is sampled at 8^2 spatial locations with $r = 16$ pixels. The contour plots represent the 50% and the 80% of the total energy of the spectral signatures.

• 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 overtakes human scale, the possible points of views become limited and more predictable. The dominant orientations on the image vary with the point of view (vanishing lines), and consequently, the spatial arrangement of the main structures (e.g., position of the ground level, horizon line).
• The *building blocks*: The building blocks refers to the elements (surfaces and objects) that compose the scene. For instance, there is a large difference between the building blocks (their shape, texture, color, etc.) of natural and man-made environments, and also between indoor and outdoor environments. The building blocks of the scene also differ from one spatial scale to another due to functional constraints and to the physical phenomena that shape the space at each scale.

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

A. 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 distance 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 (eq. 2). It is interesting to make the distinction between man-made and natural structures. We define S as the set of pictures sharing the same mean distance (D) from the observer. We study first man-made structures. The spectral signature is:

$$\overline{A}_{D,art}(f_x, f_y) = E\left[\, A(f_x, f_y) \,|\, D,\, man - made \,\right] \tag{7}$$

Figure 6 shows the spectral signatures for different mean depths. The spectral signatures can be modeled by: $\overline{A}_D(f, \theta) \sim \Gamma_D(\theta)/f^{\alpha_D(\theta)}$, as proposed in [18], where f, θ are the spatial frequencies in polar coordinates. $\Gamma_D(\theta)$ is a magnitude prefactor that is a function on orientation. The spectral signature has a linear decaying rate in logarithmic units with a slope given by $-\alpha_D(\theta)$ ([18], [23]).

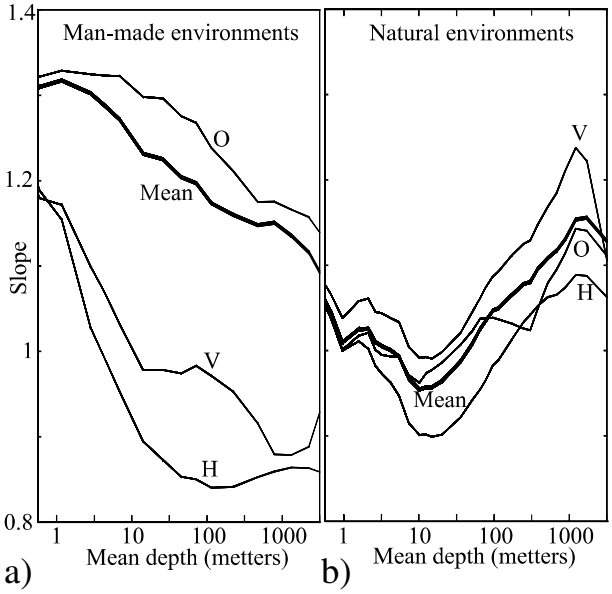

Fig. 5. Evolution of the slope of the global magnitude spectrum of real-world pictures with respect to the mean depth of the scene. The picture shows the evolution of the slope $\alpha_D(\theta)$ at 0 (Horizontal), 45 (Oblique) and 90 (Vertical) degrees, and its mean value averaged for all the orientations.

These two functions can be obtained by a linear fitting at each orientation of the spectral signature in logarithmic units (see [23] for details).

In order to illustrate the evolution of the spectral signature figure 5.a 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 [21]. 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, we observe a monotonic decreasing slope (e.g. 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 (e.g. 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. 6).

Although the increase of roughness with distance appears as something intuitive, it is not a general rule and it cannot be applied to any kind of picture. For pictures of natural environments, the spectral signature has a completely opposite behavior with respect to the mean depth (see figures 5.b and 7): 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 Figure 7 illustrate this point. For natural scene pictures, in 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. 6 and 7). To illustrate this point, figure 5 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. In average, close-up views of objects have no strong dominant orientations and thus, an isotropic amplitude spectrum.

Fig. 6. Evolution of the global spectral signature of man-made structures with respect to the mean depth. Each signature is obtained from averaging over more than 300 pictures. The contour plots represent the 50% and the 80% of the energy of the spectral signatures.

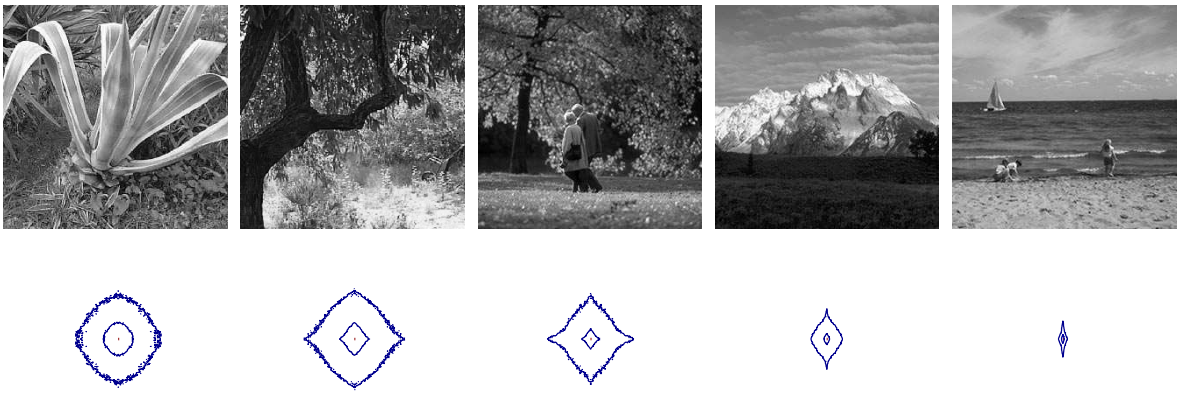

Fig. 7. Evolution of the global spectral signature of pictures of natural environments with respect to the mean depth. The contour plots represent the 50% and the 80% of the energy of the spectral signatures.

B. Relationship between the local spectral signatures and depth

Figures 8 and 9 show several examples of local amplitude spectra for man-made and natural scenes. As in equation (7), we can study the mean local amplitude spectrum (local spectral signatures) for different depths. Figure 10 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 (the variations are mostly from top to bottom). The mean 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 <10m), the spectral signatures are stationary and there is almost no bias towards horizontal and vertical orientations.
- For mean distances (50m to 500m) the spectral signatures are non-stationary 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.

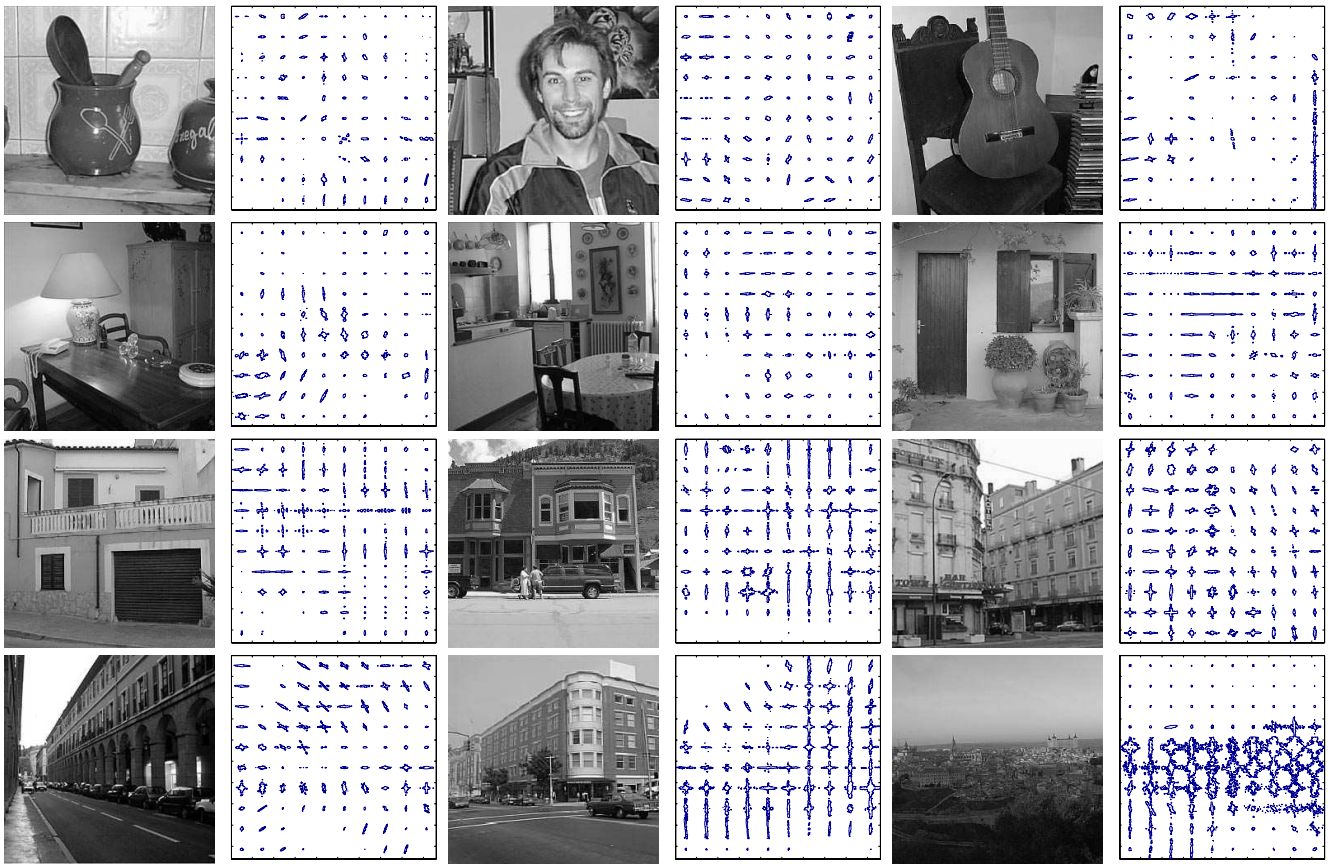

Fig. 8. Sections of the windowed Fourier transform for scenes pictures with man-made structures with increasing mean depth.

C. Discussion

To summarize, there exists a strong relationship between the structures present in the image and the mean depth of the scene (also related with the volume and scale of the space that composes the scene). This point is demonstrated in the rest of the paper by showing that absolute depth can be estimated from structural features (output energies of multiscale oriented filters).

In order to compute depth related information, several studies have used structural information. Pentland [21] mentioned the possibility of computing perspective gradients in a monocular view, by using the gradient of the Fractal dimension computed locally in the picture. Keller and collaborators [13] used the fractal dimension to differentiate between specific landscapes and natural textures like tree lines and mountain lines, and proposed to use the average Holder constant [13] for detecting scale changes. Both studies refer to relative depth measurements. There are many other studies that compute relative depth measurements using spectral features (e.g. shape from texture gradients [27]) but they are all restricted to the presence of textures with precise properties and they do not provide an absolute depth measurement.

IV. Holistic scene representation

This section is devoted to describe a simple representation of the scene structure that will be used in section V for absolute depth estimation. Scene representations are generally defined as the representation of a collection of objects and their spatial organization. Those representations require applying segmentation procedures or object recognition algorithms. However, as discussed in the precedent section, scenes belonging to the same category share common features as isolated objects do. Therefore, it is possible, in principle, to propose scene representations that bypass the analysis of individual objects

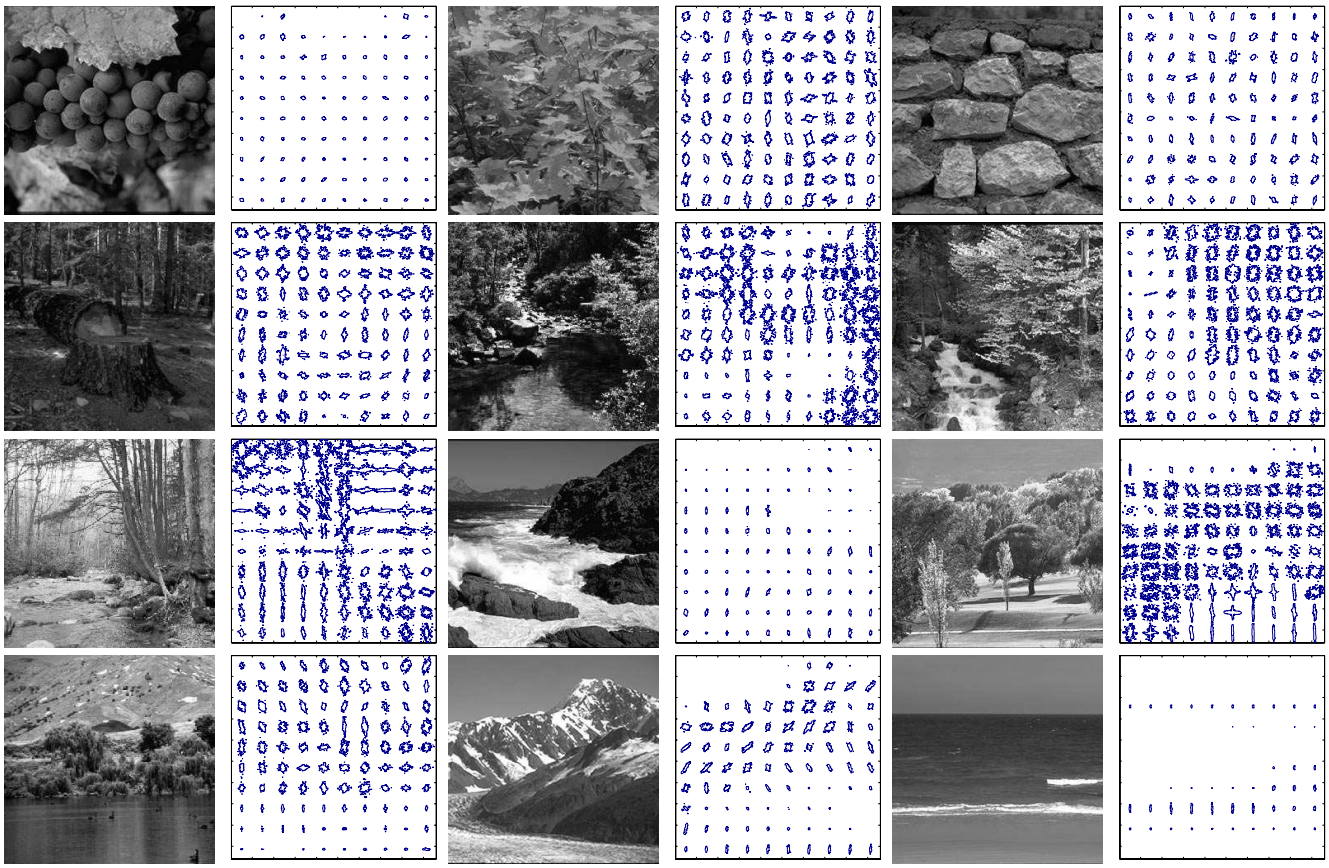

Fig. 9. Sections of the windowed Fourier transform for natural scene pictures with increasing mean depth.

by encoding the scene structure holistically. This section is devoted to a holistic representation of the scene structure (in terms of spectral components, a scene sketch) that does not require the analysis of isolated regions or objects as a first stage [18].

A. The scene sketch

A *sketch* is a simple representation of an image that captures what is essential to convey its meaning. The sketch must conserve the skeleton of the image and has to be designed with few lines fleshed out with basic textures. In that regard, drawing a sketch of an image is similar to a *reconstruction* procedure. The sketch must provide the identity of the picture and preserve 3D spatial properties (e.g. depth, perspective) of the original scene. To do so, it is not necessary to represent intrinsic object or detailed region information. As in the case of a drawing, a relevant sketch representation for machine vision must follow these constraints: 1) Low dimensionality: the sketch should be drawn from a small number of features. 2) Relevance: the representation should capture the lines and the textures that preserve the semantic identity of the scene. 3) Mapping of global semantic similarities: scenes belonging to the same semantic category should be close together in the space defined by the scene sketch features. The next sections are dedicated to the definition of a procedure that provides a sketch-like representation of complex scene pictures.

In the computational domain, robust reconstruction algorithms have been proposed for textures and homogeneous patterns and they offer an astonishing accurate duplicate of the original texture (e.g. [22], [6]). But none of these procedures is suitable for catching the non-stationary structure that the arrangement of objects and regions confer to a real world scene.

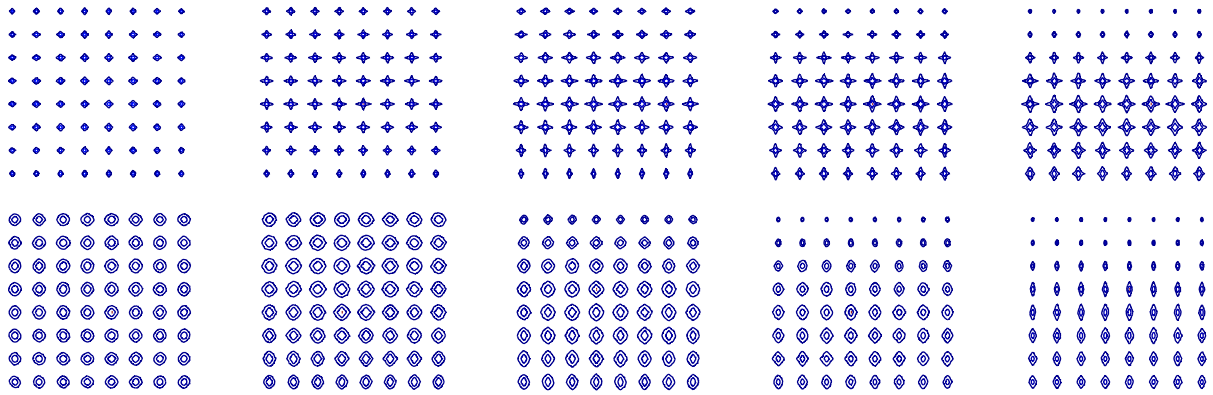

Fig. 10. Local spectral signatures (at 64 different spatial positions) for scene views of man-made (top) and natural (bottom) structures in the range (from left to right) of 1, 10, 100, 1000 meters and infinity (e.g. panoramic views ($\simeq$ 10 Km).

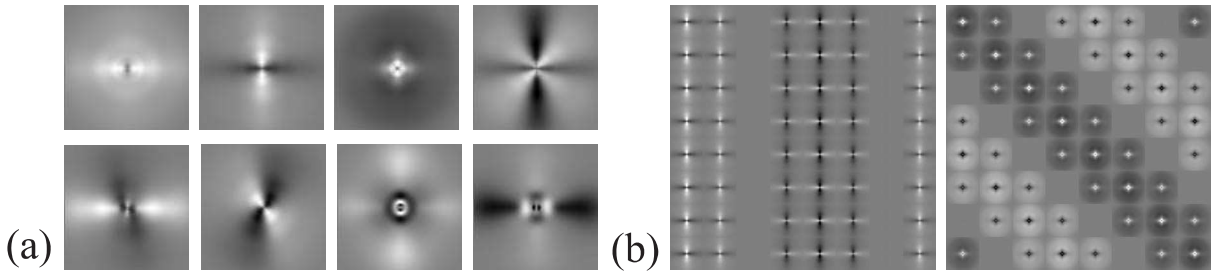

Fig. 11. a) spectral principal components, the templates correspond to $\sum_k \varphi_n(k) G_k(f_x, f_y)$. b) Examples of functions $\psi_n(\vec{x}, k)$. The figure shows how the spectral components are weighted at each spatial location to obtain a_n.

B. Low dimensional structural representation

One of the most popular features for encoding images is the output of oriented band-pass filters:

$$v(\vec{x}, k) = \sum_{\vec{x}'} i(\vec{x}') g_k(\vec{x} - \vec{x}') \tag{8}$$

where $i(\vec{x})$ is the input image and $g_k(\vec{x})$ are Gabor filters defined by $g_k(\vec{x}) = e^{\|\vec{x}\|^2/\sigma_k^2} e^{2\pi j < \vec{f_k}, \vec{x} >}$. In such a representation, $v(\vec{x}, k)$ is the output at the location $\vec{x}$ of a complex Gabor filter tuned to the spatial frequency $\vec{f_k}$. The variable k indexes filters tuned to different spatial frequencies and orientations (we use 5 bands and 8 orientations per band). Other decompositions can be used [26]. The magnitude of the filter outputs, $|v(\vec{x}, k)|$, provides contrast invariant information about the local orientation, scale and texture in the image. These features have been shown to be relevant for scene recognition tasks [18]. However, the representation achieved by $|v(\vec{x}, k)|$ is very high dimensional. In order to reduce the dimensionality we decompose the image features into its principal components:

$$a_n = \sum_{\vec{x}} \sum_{k} |v(\vec{x}, k)| \; \psi_n(\vec{x}, k) \quad with \quad |\hat{v}(\vec{x}, k)| \simeq \sum_{n=1}^{N} a_n \psi_n(\vec{x}, k) \tag{9}$$

where a_n are the decomposition coefficients and $\psi_n(\vec{x}, k)$ are the principal components of the features vector defined by $|v(\vec{x}, k)|$. Using a reduced number of coefficients a_n provides a low-resolution reconstruction (both in the spatial $\vec{x}$ and the spectral k domains) of the image features $|v(\vec{x}, k)|$. The principal components $\psi_n(\vec{x}, k)$ can also be approximated by:

$$\psi_n(\vec{x}, k) = \cos\left(2\pi \vec{x} \frac{\vec{m}_n}{L} + O_n\right) \varphi_n(k) \tag{10}$$

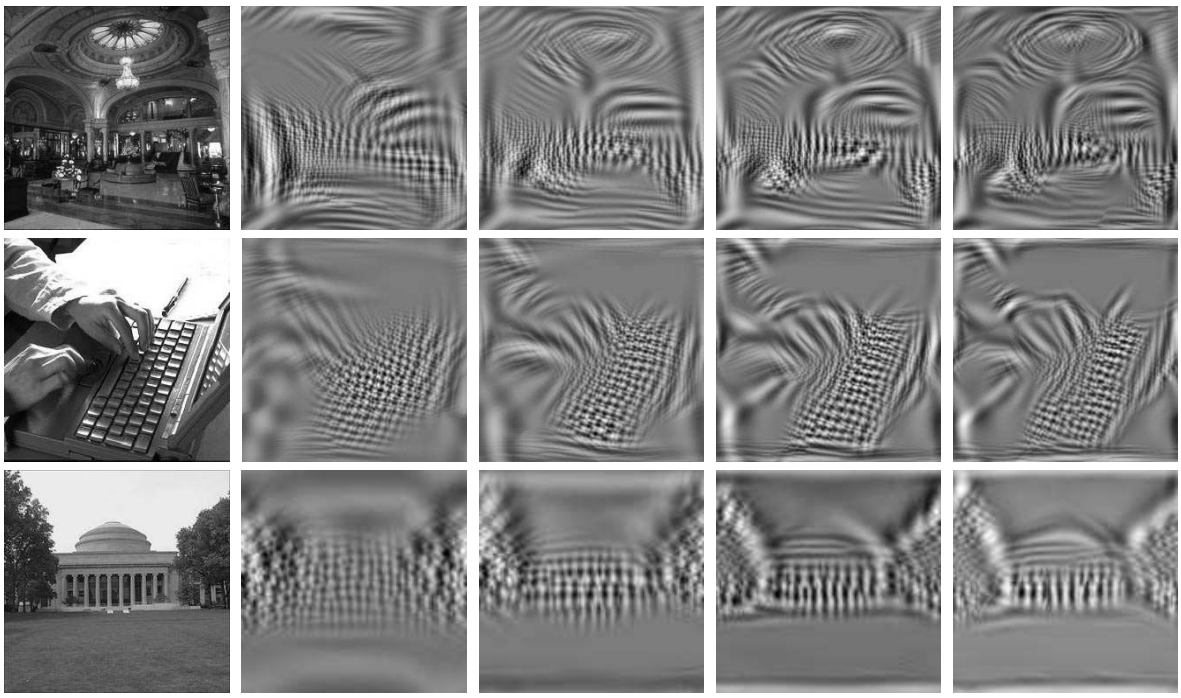

Fig. 12. Scene sketches obtained at 1, 2, 3 and 4 c/i and 12 spectral components. From left to right, the representation includes increasing spatial resolution (with constant spectral resolution). Note that small details in the original image (like the keyboard or the trees) are represented even at 1 c/i although the different spatial structures in the image are poorly localized.

where the first factor is a spatial Fourier component with $O_n = 0$ or $\pi/2$ and frequency $\vec{m}_n$ (L is the image size), and the second factor is given by the principal components of the distribution of the features vector $|v(\vec{x},k)|$ with respect to k (see Fig. 11). The interest of this approximation of the principal components is that now the functions $\psi_n(\vec{x},k)$ provide a way of controlling the amount of spatial and spectral resolution included in the representation $\vec{a} = \{a_n\}_{n=1,N}$. To summarize, $\vec{a}$ provides a holistic low-dimensional representation of the structures present in the image with a coarse description of their spatial distribution. In the next section we show that these features can be used to build a sketch of the scene.

C. *The sketch of the scene*

By an adequate set of filters, the original image $i(\vec{x})$ can be written as:

$$i(\vec{x}) \simeq \sum_k \Re eal\{v(\vec{x},k)\} = \sum_k |v(\vec{x},k)| \cos\Theta(\vec{x},k) \tag{11}$$

where $\Theta(\vec{x},k)$ is the phase of the complex value $v(\vec{x},k)$. But we do not want to deal with the high dimensional vector $v(\vec{x},k)$. In fact, we do not look for a perfect visual reconstruction as in [22], only a reconstruction that preserves the identity of the scene, i.e. a sketch of the scene. Therefore, we reconstruct the image based only in the low dimensional features vector $\vec{a}$, i.e. $|\hat{v}(\vec{x},k)|$ as given by eq. (9). The reconstruction procedure requires the estimation of the phase function $\Theta(\vec{x},k)$ that is not contained in the representation. We propose the next iterative procedure:

$$sketch^{t+1}(\vec{x}) \simeq \sum_k |\hat{v}(\vec{x},k)| \cos\hat{\Theta}^t(\vec{x},k)$$

$$\hat{\Theta}^{t+1}(\vec{x},k) = angle\left\{\sum_{\vec{x}'} sketch^{t+1}(\vec{x}')\, g_k(\vec{x}-\vec{x}')\right\} \tag{12}$$

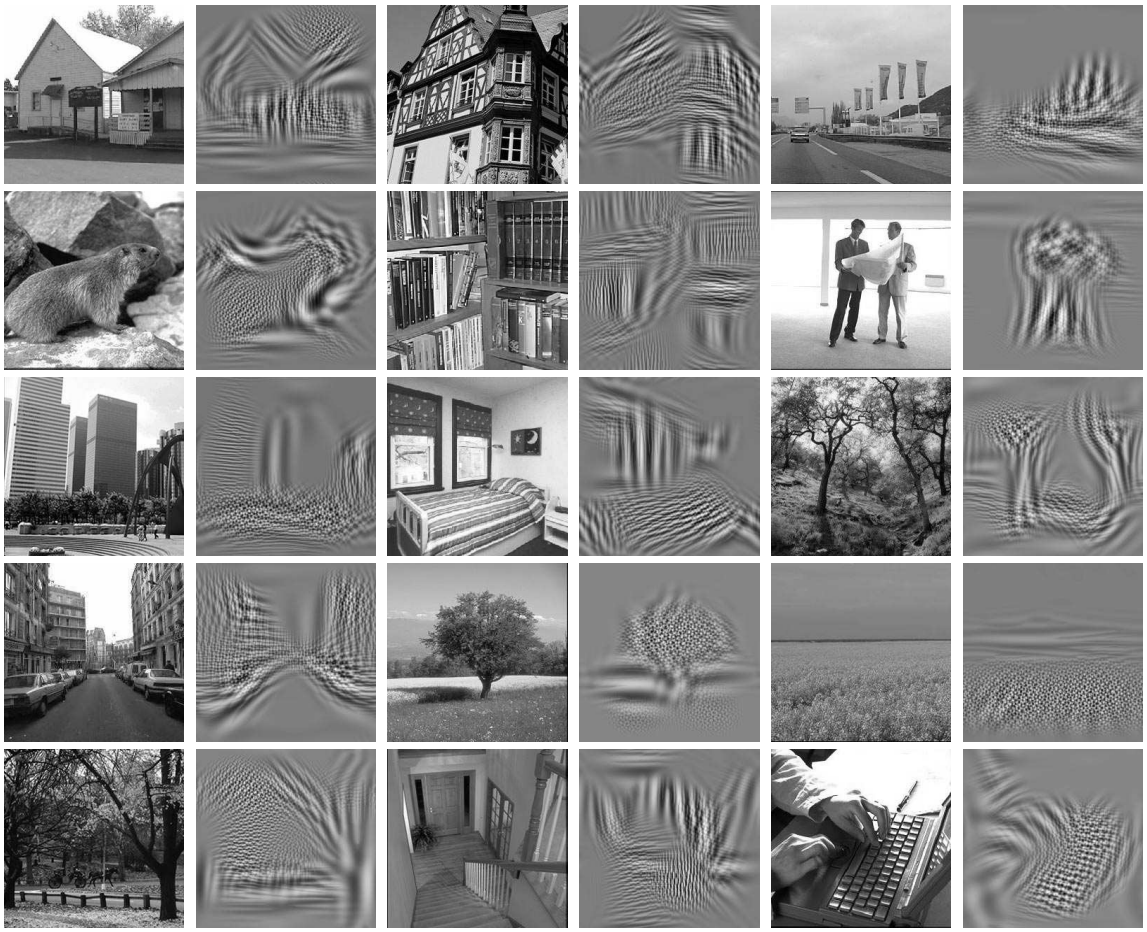

Fig. 13. Examples of scene sketches at 2 c/i.

For all the images tested, the algorithm converges to a stable image $sketch(\vec{x})$ after 20 iterations. The degree of details of the final image is determined by the number of features used in the representation $\vec{a}$. The spatial resolution can be controlled by including higher spatial frequencies in the functions $\psi_n(\vec{x}, k)$ as shown in eq. (14). Fig. 12 shows the reconstructions obtained with spatial resolutions at 1, 2, 3 and 4 cycles/image. The resulting images are sketches of the original picture and reveal the dominant edges and textures in the image. As resolution increases, the sketch includes more lines and textures. It is interesting to see how, even at 2 c/i, the resulting sketch already conveys some of the important lines that configure the space and the dominant shapes of the original scene (see Fig. 13).

As the scene sketch provides a coarse representation of the major lines in the scene, a similarity measure based on the features of the scene sketch will account for the similarity in the global structure of two scenes. In order to illustrate this point, fig. 14 shows the distribution of a set of pictures of outdoor environmental scenes organized with respect to the first two principal components of the scene sketch features. As expected, pictures are organized according to global spatial similarities: from left to right, the space differentiates among open environments and enclosed scenes. From top to bottom, the space discriminates between man-made and natural environments.

To summarize, the scene sketch features provide a holistic low-dimensional representation in which the scene is encoded as a single unit and not as the juxtaposition of objects or pre-segmented textured regions. Therefore, it provides scene recognition before segmentation.

In the next section we show how the holistic scene encoding can be used for estimating the absolute depth of the scene.

Fig. 14. Organization of scenes according to the two first components obtained by applying PCA of the sketch features. A coarse organization of scenes emerges: man-made vs. natural environments and panoramic vs. enclosed environments. It allows a first discrimination between man-made and natural environments (83%) and open versus enclosed environments (88%).

V. Absolute depth estimation

We refer to the mean depth of a scene as a measurement of the mean distance between the observer and the main components of the scene. In the absence of binocular information (or other direct sources of information), absolute depth measurements must rely on knowledge about the world. Although, the relative 3D structure of the scene can be estimated by using pictorial cues as perspective gradients, edges, occlusions, etc. the estimation of the absolute size of the elements in the scene might require access to knowledge about the world (figure 15). For instance, absolute depth can be measured based on the scale of familiar objects with known size. But this strategy needs a representation of the objects in the scene.

In this section, we propose to use the scene-centered representation detailed in section IV in order to estimate absolute depth. In such an approach, the depth of the scene is not given by the recognition

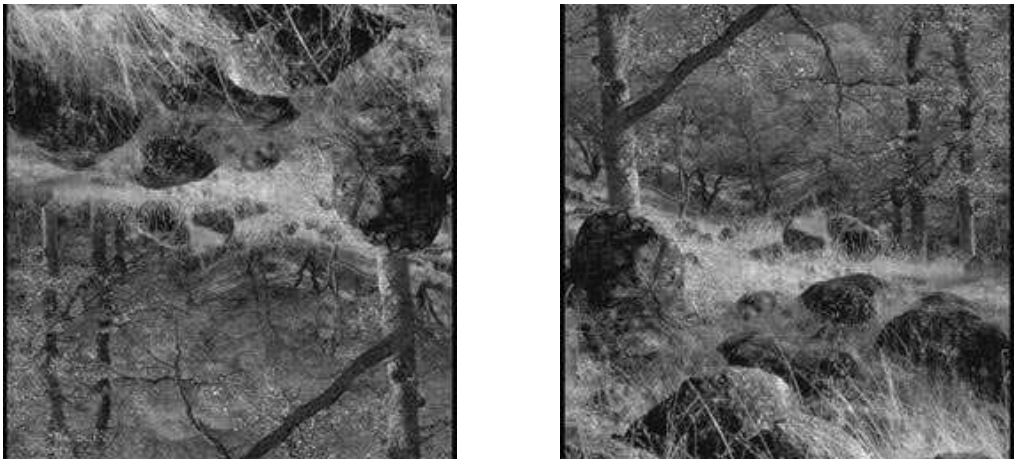

Fig. 15. Effects of recognition on depth judgments. The image on the left is generally recognized as close-up view on bushes and maybe a spider grid on top. The right-hand side, the image is categorized as the inside of a forest corresponding to a larger distance than the image on the left. The image on the left corresponds to the image on the right after inversion upside-down. The inversion affects the perception of concavities and convexities due to the assumption of light from above, and, therefore, modifies the perceived relative 3D structure of the scene. But, moreover, the wrong recognition affects the absolute scale of the perceived space.

of individual objects but by the recognition of the whole scene as a unique entity. As an illustration, from Fig. 14 we can see that scenes sharing similar depths are close together in the features space defined in the precedent section.

To the contrary of 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 (in terms of spectral features) and the mean depth of the scene. 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 (a mountain) and that the texture introduced by trees belongs to a medium-scale structure. In this section, we present a probabilistic framework for modeling the relationship between depth and the scene structure using the features provided by the sketch representation. Note that the system does not know that the landscape contains mountains, trees or grass. None object information is never given at any point.

The problem of depth estimation from the scene features can be formulated as a regression problem. Given a training set of T images with features $\{\vec{a}_i\}_{i=1,T}$ and depth $\{d_i\}_{i=1,T}$ we want to obtain a function $D = f(\vec{a})$. A simple estimator can be built from the conditional PDF $p(D|\vec{a})$. We estimate the PDF using the Parzen-window:

$$p(D|\vec{a}) = \frac{\sum_{i=1}^{T} K_{\sigma_1}(D - d_i) K_{\sigma_2}(\vec{a} - \vec{a}_i)}{\sum_{i=1}^{T} K_{\sigma_2}(\vec{a} - \vec{a}_i)} \tag{13}$$

where the kernel $K_\sigma(\vec{x})$ is a gaussian kernel with width σ.

The training database used for the results presented here is composed of more than 4000 pictures for which the two authors reported the absolute distance of the original pictures. The image database covers a large variety of outdoor and indoor scenes and man-made and natural environments with a large range of distances (from less than a meter to kilometers) as well as close-up views of objects and natural textures. We used the reported depth of each image for training the system.

Given a new scene picture, the mean depth is estimated from the scene sketch features as $\overline{D} =$

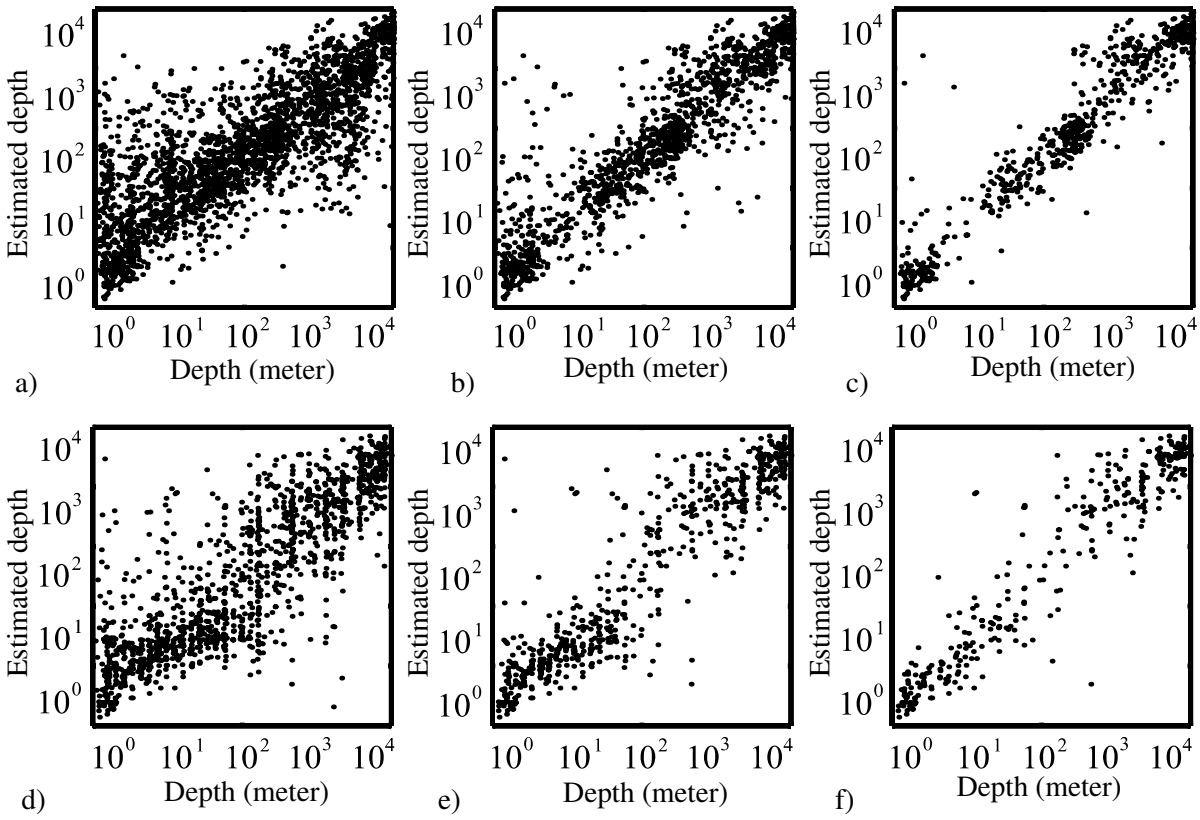

Fig. 16. Absolute depth estimation results (units are in meters) for both man-made (a, b and c) and natural (d, e and f) environments. From left to right we reduce the number of images in which the confidence level is high enough to provide a depth estimate. At the left-hand side (a and d) we show the estimation results for all the images from the test set. In the center (b and e) only the 50% of the images that have the highest confidence levels are shown, and at the right-hand side (c and f), only the 25% of the images from the test set with the highest confidence levels are evaluated.

$E[D \mid \vec{a}]$:

$$\overline{D} = \frac{\sum_{i=1}^{T} d_i K_{\sigma_2}(\vec{a} - \vec{a}_i)}{\sum_{i=1}^{T} K_{\sigma_2}(\vec{a} - \vec{a}_i)} \tag{14}$$

The test was also performed using a large variety of scenes. Fig. 16.a shows the comparison between the real depth reported by the subjects and the estimated depth. Fig. 17 shows 7 pictures of man-made environments sorted according to the mean depth estimated by eq. (14). On average, the organizations provided by the algorithm have a spearman rank correlation of 0.82 with respect to the orderings of the same pictures performed by subjects. The estimation of the PDF $p(D|\vec{a})$ also provides a method to measure the reliability of the estimation provided by eq. (14) for each new picture:

$$\overline{\sigma_D^2} = \frac{\sum_{i=1}^{T} (\sigma_1^2 + d_i^2) K_{\sigma_2}(\vec{a} - \vec{a}_i)}{\sum_{i=1}^{T} K_{\sigma_2}(\vec{a} - \vec{a}_i)} - \overline{D}^2 \tag{15}$$

For each new image, the bigger is the value of the variance $\overline{\sigma_D^2}$ the less reliable is the estimation. Therefore, if provides a way to decide when we can reliably assign the estimated depth to an image. Fig. 16 shows the performances for both natural and man-made structures for the entire test database and also when depth is assigned to the 50% of images with the highest confidence level (fig. 16.b and e) and when depth is assigned to the 25% of images with the highest confidence level (fig. 16.c and f). For the 50% of the images selected, the correlation between the estimated depth and the depth reported by the subjects is 0.92. The results show that the variance estimation provides an effective

Fig. 17. Example of scenes organized with respect to the estimated mean depth.

way of predicting the reliability of the estimation. Figures 18-21 show examples of pictures of man-made and natural scenes sorted according to the estimated depth and the corresponding conditional probability $p(D|\vec{a})$ estimated from the scene sketch of each picture. Figures 18 and 20 show images for which the function $p(D|\vec{a})$ is narrow having one unique mode. Those images produce high confidence ratings and therefore, there are few errors in the estimated depths. Figures 19 and 21 show images selected among the ones with the largest variance estimations. The corresponding conditional PDFs have multiple modes covering a large range of possible depths. Those images have low confidence and there are many errors in the estimated depths.

There are several potential applications for a global measure of absolute mean depth. For instance, adding the estimation of the mean depth of a scene to other attributes (like color, orientation, texture) may significantly increase performances of semantic recognition in applications like image indexing. Figure 22 shows the distribution, along the mean depth axis, of basic scene categories commonly employed by human observers when asked to name images. 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.

VI. Task and context-dependent automatic 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 pre-processing 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 [14], [15] 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 [30]. This provides 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. The expected size can be approximated by $\widetilde{S}_{obj} \simeq K_{obj}/Depth^2$, were K_{obj} is a normalization constant giving the size of the object at 1 meter. This will give a restriction of the possible scales as $Depth$ here refers to the mean depth. Figure 23 illustrates how the procedure can predict the sizes of people and cars for different environments computed using the mean depth. If one object has an expected size larger than the volume of the space or it is smaller than the pixel size then it can be discarded from the search.

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. Height is almost invariant to heads pose (most of pose variations are due to rotation in an horizontal plane). Using logarithmic units there is a linear dependence between heads height and distance: $log(H) = k - log(D)$, where H is the true height measured directly from the image. For the 83% 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. As a consequence, the estimated distance ($D = K/H$) is also in the interval $[D/2, D*2]$ for

Fig. 18. Example of scenes organized with respect to the estimated mean depth. For each image, the figure shows also the shape of the function $p(D|\vec{a})$. The horizontal axis (depth) is in a logarithmic scale. The examples shown in this figure produce narrow conditional probabilities and therefore are assigned high confidence.

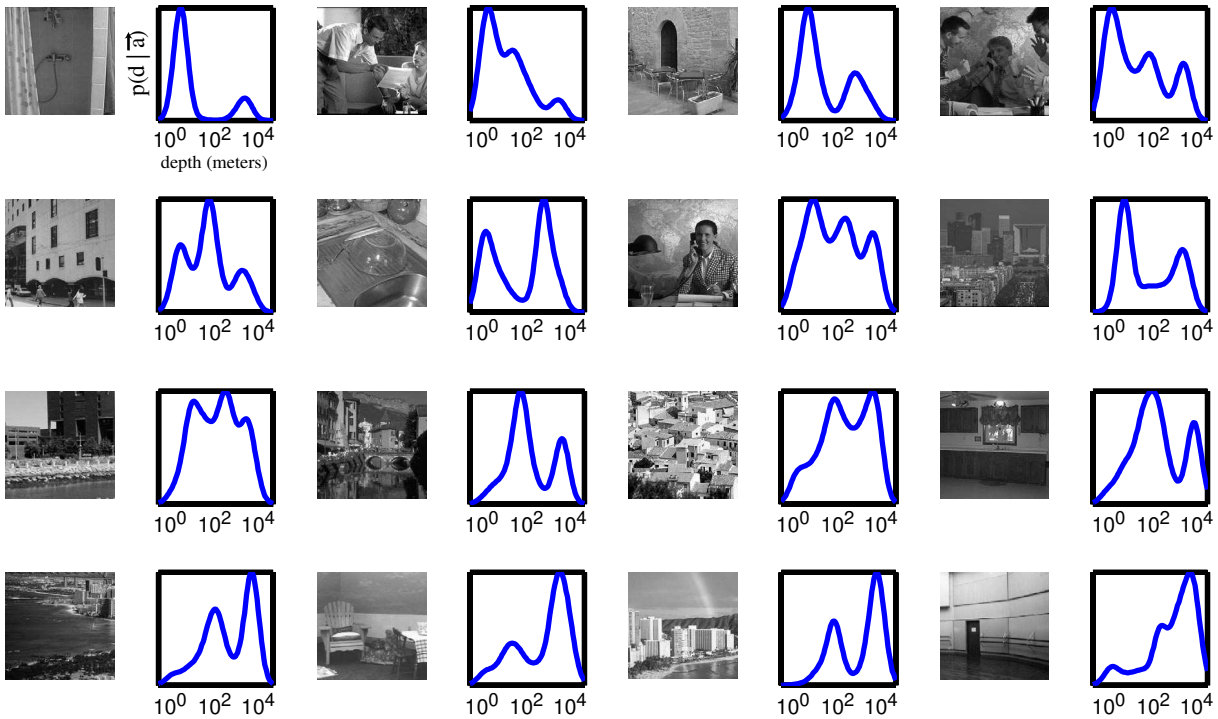

Fig. 19. Example of scenes organized with respect to the estimated mean depth. These examples produce wide conditional probabilities with multiple modes and therefore are assigned low confidence. Many of these images are assigned an incorrect mean depth.

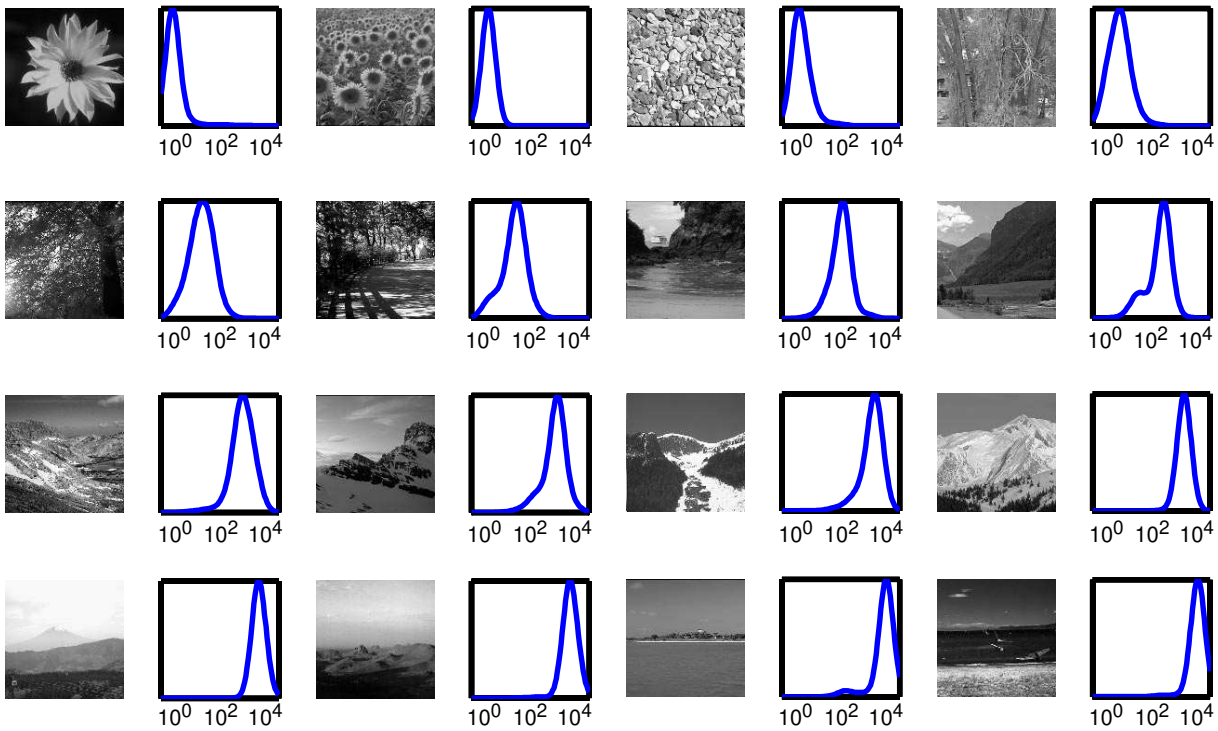

Fig. 20. Example of scenes organized with respect to the estimated mean depth. These examples produce narrow conditional probabilities and therefore are assigned high confidence.

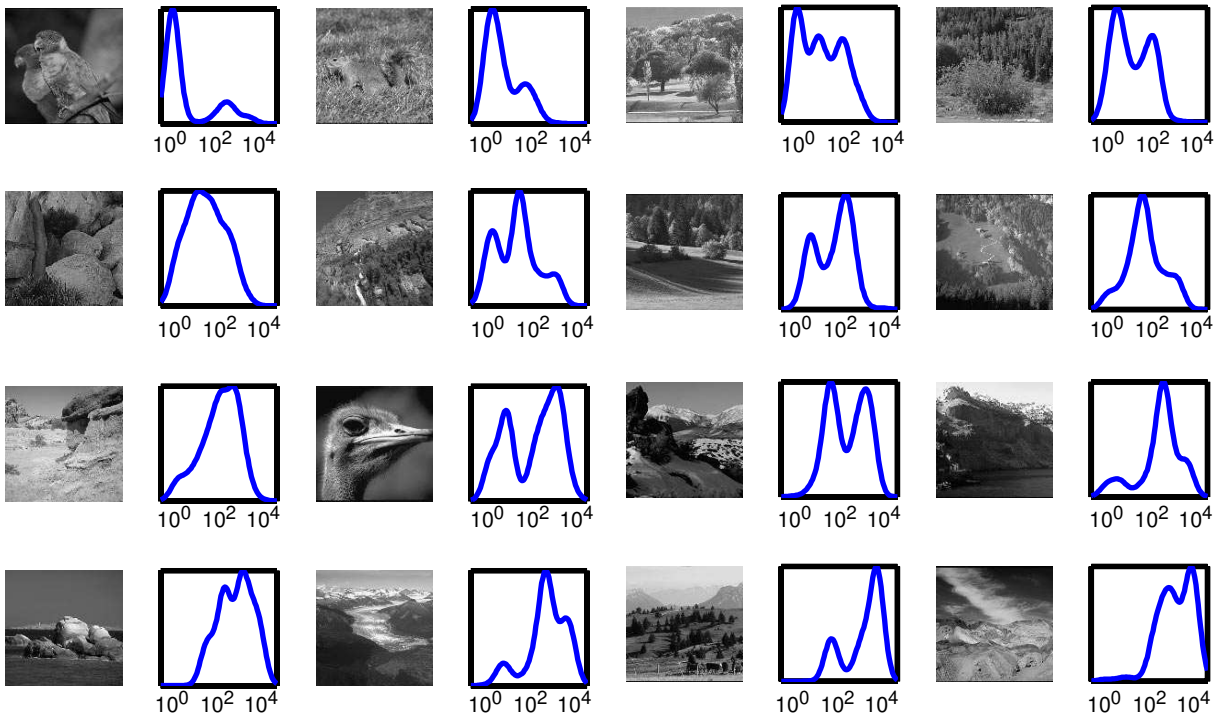

Fig. 21. Example of scenes organized with respect to the estimated mean depth. These examples produce wide conditional probabilities with multiple modes and therefore are assigned low confidence. Many of these images are assigned an incorrect mean depth.

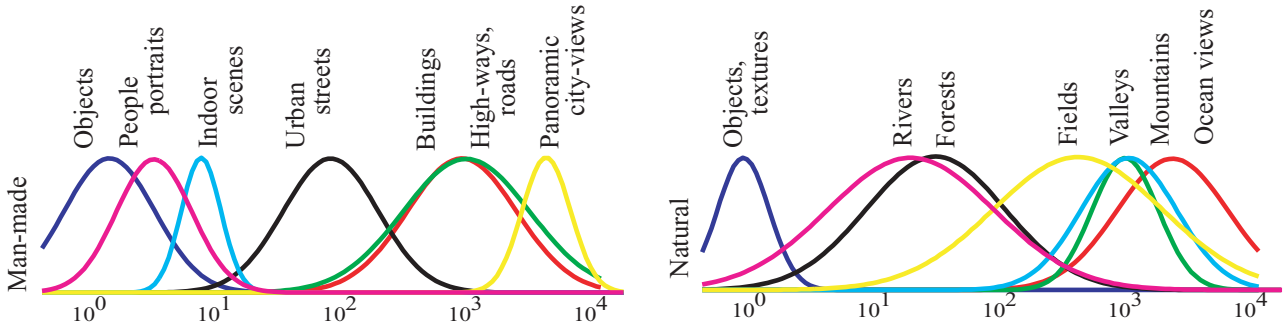

Fig. 22. Distribution of scene categories commonly used by subjects in function of the mean depth of the scene for man-made and natural structures. This distributions are obtained from a separate annotation for another study [18] of the images database used for the depth estimation. The distributions are then obtained by fitting a gaussian distribution to the PDF $p(D|category)$ obtained for each scene category.

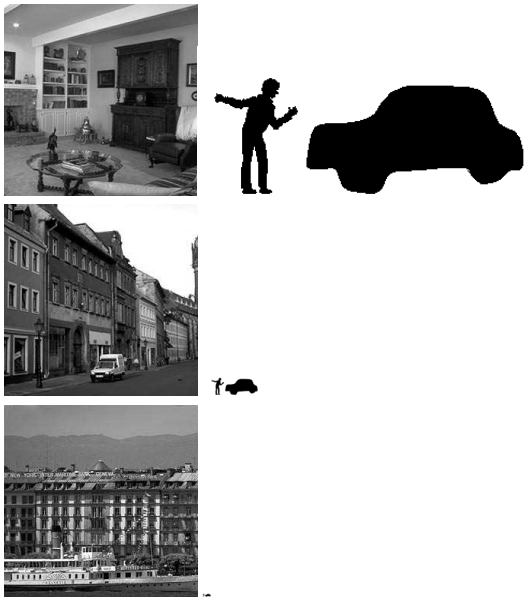

Fig. 23. Task and context-dependent scale selection. The schematic drawings represent the attended size of people and cars estimated using the mean depth as descriptor of the context information.

the 83% of scenes tested. Note that the estimated distance (D) of people in the pictures is obtained without any detection stage of faces or bodies but by using the whole picture as the context in which heads are located [18], [30].

VII. Conclusion

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 allow a estimation of depth by recognizing the structures present in the scene.

Depth computation as proposed here does not require recovering the local 3D structure of the scene as an intermediate step. Rather simple image parameters provide absolute depth related information that does not require object recognition, processing of surfaces, shadows, or junctions. Therefore, 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.

Mean depth is a capital attribute for recognizing the scene in front of the observer. Combined with other perceptual attributes of the space that the scene picture represents (open, large, bounded, etc. See [18]) it can allow the recognition of the semantic category of the scene as a first step in the visual processing before the analysis of edges, surfaces or object detection.

References

[1] Baddeley, R. 1997. The correlational structure of natural images and the calibration of spatial representations. *Cognitive science* **21**, 351–372.

[2] Barrow, H. G., and Tenenbaum, J. M. 1981. Interpreting line drawings as tree-dimensional surfaces. *Artificial intelligence* **17**, 75–116.

[3] Bergen, J. R., and Landy, M. S. 1991. Computational Modeling of Visual Texture Segregation. In *Computational Models of Visual Processing*. M. S. Landy and J. A. Movshon (Eds.). pp. 253–271, Cambridge, MA: MIT Press (1991).

[4] Carson, C., Belongie, S., Greenspan, H., and Malik, J. 1997. Region-based image querying. *Proceedings IEEE Workshop on Content-Based Access of Image and Video Libraries*, pp: 42–49.

[5] De Bonet, J. S., and Viola, P. 1997. Structure driven image database retrieval. *Advances in Neural Information Processing* **10**.

[6] Efros, A., and Freeman, W. T. 2001. Image quilting for texture synthesis and transfer. *SIGGRAPH 2001*.

[7] Field, D. J. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America*, 4, 2379–2394

[8] Gorkani, M. M., and Picard, R. W. 1994. Texture orientation for sorting photos "at a glance". *Proc. Int. Conf. Pat. Rec.*, Jerusalem, Vol. I, 459–464.

[9] Hancock, P. J., Baddeley, R. J., and Smith, L. S. 1992. The principal components of natural images. *Network* 3, 61–70.

[10] Horn, B. K. P., and Brooks, M. J. 1989. *Shape from shading*. MIT Press, Cambridge, MA.

[11] Jepson, A. Richards, W., and Knill, D. 1996. Modal structures and reliable inference. *Perception as Bayesian Inference*, eds. D. Knill and W. Richards, Cambridge Univ. Press, pp. 63-92.

[12] Jordan, M. I., and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6:181–214.

[13] Keller, J. M., Crownover, R. M., and Chen, R. Y. 1987. Characteristics of natural scenes related to the fractal dimension. *IEEE transactions on pattern analysis and machine intelligence*, 9(5):621–627

[14] T. Lindeberg. 1993. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *International Journal of Computer Vision*, 11(3):283–318.

[15] T. Lindeberg. 1998. Principles for automatic scale selection. *International Journal of Computer Vision*, 30(2):77–116.

[16] Liu, F., and Picard, R. W. 1996. Periodicity, directionality and randomness: Wold features for image modeling and retrieval. *IEEE transactions on Pattern Analysis and Machine Intelligence*. 18:722-733

[17] Oliva, A., Torralba, A., Guerin-Dugue, A., and Herault, J. 1999. Global semantic classification using power spectrum templates. *Proceedings of The Challenge of Image Retrieval*. Electronic Workshops in Computing series, Springer-Verlag. Newcastle.

[18] Oliva, A., and Torralba, A. 2001. Modeling the Shape of the Scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145-175.

[19] Palmer, S. E. 1999. *Vision Science*. MIT Press, Cambridge, MA.

[20] Papoulis, A. *Probability, random variables and stochastic processes*. MacGraw-Hill, second edition, 1984.

[21] Pentland, A. P. 1984. Fractal-based description of natural scenes. *IEEE transactions on pattern analysis and machine intelligence*, 6(6):661–674

[22] Portilla, J., and Simoncelli, E.P. 2000. A parametric texture model based on joint statistics of complex wavelets coefficients. *International Journal of Computer Vision*, 40:49-71.

[23] van der Schaaf, A., and van Hateren, J.H. 1996. Modeling of the power spectra of natural images: statistics and information. Vision Research, 36, 2759–2770.

[24] Schyns, P. G., and Oliva, A. 1994. From blobs to boundary edges: evidence for time- and spatial- scale dependent scene recognition. *Psychological Science*. 5:195-200

[25] Shimshoni, I., Moses, Y., and Lindenbaum, M. 2000. Shape reconstruction of 3D bilaterally symmetric surfaces. *International Journal of computer vision*, 2:1–15

[26] Simoncelli, E. P., and Freeman, W. T. 1995. The Steerable Pyramid: A flexible architecture for multi-scale derivative computation. *2nd IEEE Int'l Conf on Image Processing*, Washington DC, October 1995.

[27] Super, B. J., and Bovik, A. C. 1995. Shape from texture using local spectral moments. *IEEE transactions on pattern analysis and machine intelligence*, 17(4):333–343

[28] Szummer, M., and Picard, R. W. Indoor-outdoor image classification. In *IEEE intl. workshop on Content-based Access of Image and Video Databases*, 1998.

[29] Torralba, A., and Oliva, A. 1999. Scene organization using discriminant structural templates. *Proc. Of Int. Conf in Comp. Vision*, ICCV99, 1253-1258.

[30] Torralba, A., and Sinha, P. Statistical context priming for object detection: scale selection and focus of attention. 2001. AI-Memo 2001-020, CBCL Memo 205, September.

[31] Vailaya, A., Jain, A., and Zhang, H. J. 1998. On image classification: city images vs. landscapes. *Pattern Recognition*, 31:1921–1935

[32] Vailaya, A., Figueiredo, M., Jain, A., and Zhang, H. J. 1999. Content-based hierarchical classification of vacation images. *Proceedings of the International Conference on Multimedia, Computing and Systems*,June.