International Journal of Computer Vision 42(3), 145–175, 2001

Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope*

AUDE OLIVA
Harvard Medical School and the Brigham and Women's Hospital, 221 Longwood Ave., Boston, MA 02115
oliva@search.bwh.harvard.edu

ANTONIO TORRALBA
Department of Brain and Cognitive Sciences, MIT, 45 Carleton Street, Cambridge, MA 02139
torralba@ai.mit.edu

Received February 7, 2000; Revised January 22, 2001; Accepted January 22, 2001

Abstract. In this paper, we propose a computational model of the recognition of real world scenes that bypasses the segmentation and the processing of individual objects or regions. The procedure is based on a very low dimensional representation of the scene, that we term the *Spatial Envelope*. We propose a set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) that represent the dominant spatial structure of a scene. Then, we show that these dimensions may be reliably estimated using spectral and coarsely localized information. The model generates a multidimensional space in which scenes sharing membership in semantic categories (e.g., streets, highways, coasts) are projected closed together. The performance of the spatial envelope model shows that specific information about object shape or identity is not a requirement for scene categorization and that modeling a holistic representation of the scene informs about its probable semantic category.

Keywords: scene recognition, natural images, energy spectrum, principal components, spatial layout

I. Introduction

Seminal conceptions in computational vision (Barrow and Tannenbaum, 1978; Marr, 1982) have portrayed scene recognition as a progressive reconstruction of the input from local measurements (edges, surfaces), successively integrated into decision layers of increasing complexity. In contrast, some experimental studies have suggested that recognition of real world scenes may be initiated from the encoding of the global configuration, ignoring most of the details and object information (Biederman, 1988; Potter, 1976). Computational and experimental schools achieve different objectives of recognition: for the former, *recognition* is a reconstruction procedure of the 3D scene properties that is an essential step in tasks involving movement or grasping. For the latter, *recognition* of the scene implies providing information about the semantic category and the function of the environment.

In the research described hereafter, we propose a computational model of the recognition of scene categories that bypasses the segmentation and the processing of objects. In that regard, we estimate the structure or "shape of a scene" using a few perceptual dimensions specifically dedicated to describe spatial properties of the scene. We show that holistic spatial scene properties, termed *Spatial Envelope properties*, may be reliably estimated using spectral and coarsely localized information. The scene representation characterized by the set of spatial envelope properties provides a meaningful description of the scene picture and its semantic category.

*The authors contributed equally to this work.

The paper is organized as follows: Section II describes recent experimental results about scene recognition. Section III introduces the concept of *Spatial Envelope* as a holistic descriptor of the main structure of a scene. Section IV gives an overview of computational models of scene recognition, and explains our approach, based upon the spectral signature of scene categories. Section V addresses the computational model per se, and details the computation of the spatial envelope properties. Section VI evaluates recognition performance of the model. The last section discusses issues related to the problem of semantic scene recognition.

II. Scene Recognition

A. *What is a Scene?*

This study is dedicated to the representation of environmental scenes (see section V.B for a description of the database). In an attempt to define what a "scene" is, as opposed to an "object" or a "texture", we propose to consider the absolute distance between the observer and the fixated zone. Therein, if an image represents an "object" when the view subtends 1 to 2 meters around the observer, a "view on a scene" begins when there is actually a larger space between the observer and the fixated point, usually after 5 meters. Thus, whilst most of the "objects" are at a hand distance, a scene is mainly characterized as a place in which we can move.

B. *Scene Recognition Studies*

A number of experimental studies have demonstrated that we integrate enough information about the meaning of a scene in less than 200 ms (Potter, 1975; for a review see Henderson and Hollingworth, 1999). In fact, we recognize its "gist"[1] as quickly and as accurately as a single object (Biederman, 1988). Fast scene recognition performances may be mediated by the spatial arrangement of the objects and the scene ground plane (e.g. the "spatial layout", Hochberg, 1968; Sanocki and Epstein, 1997). Scene meaning may also be driven from the arrangement of simple volumetric forms, the "geons" (Biederman, 1987), and even by the spatial relationships between unspecified blobs of specific size and aspects ratios (Schyns and Oliva, 1994; Oliva and Schyns, 1997; see also Carson et al., 1997,1999, in the computational domain). These studies among others (see Rensink et al., 1997) suggested that object information might be spontaneously ignored during the rapid categorization of environmental pictures. Coarse blobs made of spatial frequency as low as 4 to 8-cycles/image provided enough information for instant recognition of common environments even when the shape and the identity of the objects could not be recovered (Oliva and Schyns, 2000). In a related vein, several studies about scene viewing and scrutinizing (O'Regan et al., 1999; Rensink, 1999; Rensink et al., 1997; Simon and Levy, 1997) have demonstrated that subjects can be totally blind to object changes (as displacement and suppression) even when these changes affect meaningful parts of the scene. To summarize, experimental evidence suggests that, when viewing a scene for a short time, we extract enough visual information to accurately recognize its functional and categorical properties (e.g., people in a street, surrounded by tall buildings) whereas we overlook most of the perceptual information concerning the objects and their locations. The primary semantic representation appears to be built on a low resolution spatial configuration.

III. The Spatial Envelope: A Representation of the Shape of a Scene

A. *The Shape of a Scene*

From all of the visual properties, shape is the mandatory property that carries the identity of visual stimuli. However, if theories of holistic shape representation for objects are well acknowledged (e.g., template matching), speaking about the *shape of a scene* appears odd at first. A scene is usually understood as an unconstrained configuration of objects, and consequently, its semantic recognition may need to initially find the objects and their exact location. Looking at the pictures of Fig. 1, it is almost impossible to neglect the identities of objects, therein, we have the impression of actively using them for recognizing the scene. In this paper, rather than looking at a scene as a configuration of objects, we propose to consider a scene like an individual object, with a unitary shape. As an illustration, Fig. 1 shows a surface presentation of images. Along this representation, the 3D cues are not perceived anymore neither are the precise forms nor the identity of objects. While the scene pictures and their surface representations contain the same pictorial information, the surface pictures may be conceptualized as a unitary form (rather complex), with concave and convex regions of different sizes and amplitudes.

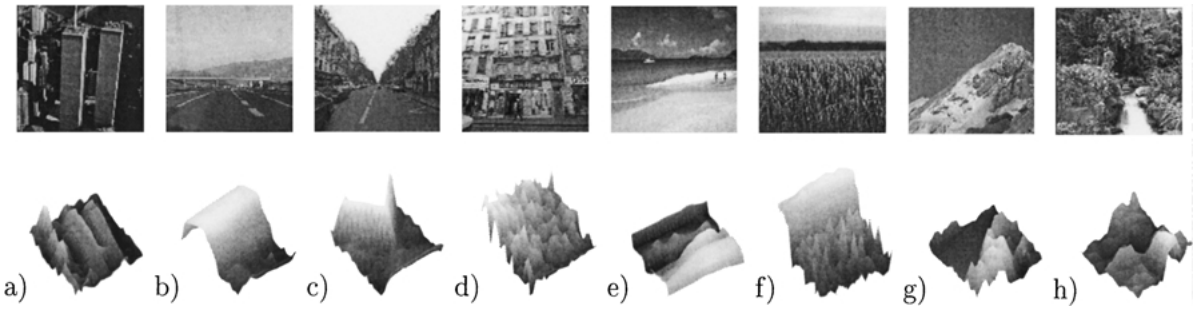

Figure 1. Scenes with different spatial envelopes and their surface representation, where the height level corresponds to the intensity at each pixel (images were low-passed): a) skyscrapers, b) an highway, c) a perspective street, d) view on a flat building, e) a beach, f) a field, g) a mountain and e) a forest. The surface shows the information really available after projection of the 3D scene onto the camera. Several aspects of the 3D scene have a direct transposition onto the 2D surface properties (e.g., roughness).

Similarly to object categories like car or animal, in which the exemplars usually look alike because they have the same "function", we will show that scenes belonging to the same category share a similar and stable spatial structure that can be extracted at once, and without segmenting the image. We will show that perceptual properties exist that can be uncovered from simple computations, and that these properties can be translated into a meaningful description of the space of the scene. We term this description the *Spatial Envelope* representation, as it mainly refers to qualities of the space (e.g., open, small).

B. Definition of the Spatial Envelope

Similar to the vocabulary employed in architecture, the *Spatial Envelope* of an environment is made by a composite set of boundaries, like walls, sections, ground, elevation and slant of the surfaces that define the shape of the space. For example, the spatial boundaries of most city views would be made of the facades of the buildings connected to the pavement as a ground and the sky as a ceiling. Most freeways look like a large surface stretching to the horizon line, filled-in with concavities (e.g., vehicles) whereas a forest scene will comprise an enclosed environment, vertically structured in the background (trees), and connected to a textured horizontal surface (grass). The spatial envelope is represented by the relationship between the outlines of the surfaces and their properties including the inner textured pattern generated by windows, trees, cars, people etc.

C. Spatial Categories

Environments are commonly named using precise semantic terms, like beach, street or forest (see Tversky and Hemenway, 1983). But that level of description does not explicitly refer to the scene structure. Therein, we performed a categorization task intended to describe the dimensions that would sketch the spatial envelope of the scene.

Several studies have proposed perceptual properties for representing texture images (Amadasun and King, 1989; Rao and Lohse, 1993; Tamura et al., 1978; Heaps and Handel, 1999). Rao and Lohse (1993) found that the three first dimensions used for distinguishing between textures are repetitiveness (vs. irregularity), contrast (correlated with directionality) and the degree of granularity or complexity of the texture. But in the context of real-world scenes, identification of such properties is an unexplored research field. Similar descriptions have often been reduced to vague notions such as orientations of contours, boundaries, clusters of "forms" (Biederman, 1988), or low spatial frequency blobs (Schyns and Oliva, 1994; Oliva and Schyns, 2000). In this respect, we designed an experiment for identifying meaningful dimensions of the scene structure.

Seventeen observers were asked to split 81 pictures into groups. They were told that scenes put in the same group should have a similar global aspect, a similar global structure or similar elements. They were explicitly told not to use a criteria related to the objects (e.g., cars vs. no cars, people vs. no people) or a scene semantic groups (e.g., street, beach). The task consisted of three steps. The first step was to divide the 81 pictures into two groups. In the second step, subjects split each of the two groups into two more subdivisions, and in the third step, subjects split the four groups into two groups each, leaving a total of 8 subgroups. At the end of each step, subjects were asked to explain the criteria they used in a few words. The taxonomy shown

Table 1. Spatial envelope properties of environmental scenes.

Property	S1	S2	S3	Total
Naturalness	65	12	0	77
Openness	6	53	24	83
Perspective	6	18	29	53
Size	0	0	47	47
Diagonal plane	0	12	29	41
Depth	18	12	29	59
Symmetry	0	0	29	29
Contrast	0	0	18	18

Results are in %, for each of the three experimental steps. The total represents the percent of times the attribute has been used regardless of the stage of the experiment.

in Table 1 summarizes the different criteria. The two first criteria concerned the *naturalness* status of the environment (man-made scenes, urban vs. natural landscapes) and the *openness* of the environment, respectively chosen by 77% and 83% of the subjects. The notion of openness was mainly described as open vs. closed-enclosed environment, scenes with horizon vs. no horizon, a vast or empty space vs. a full, filled-in space. Three other important criteria were *perspective* (mostly used for urban scenes), *size*, (referring to small or detailed vs. big elements) and *diagonal planes*. That last criteria mostly referred to undulating landscapes, mountains and rocks, and was also termed as "elevation plane", "contours going down", "falling lines", or "sharpness". In terms of depth criteria, the descriptions given by subjects were not unique, as they were referring to different space properties. Three observers who responded "depth" were actually referring to distant "open scenes" vs. proximate and "closed scenes". Two subjects who choose "depth" actually grouped images according to the degree of "expansion" of the environment (going away vs. closed) and five other subjects meant the "size" of the elements ("close and small" vs. "far and large"). Lastly, criteria of major importance for object classification such as the symmetry and the contrast were poorly chosen. It may seem surprising that symmetry does not emerge as a major spatial property, but that result is in agreement with another experimental study showing that symmetry (and also continuity) are not constraints taken into account when subjects have to quickly recognize an environmental scene (Sanocki and Reynolds, 2000).

D. Spatial Envelope Properties

Based on the experimental results, we considered the following five spatial envelope properties.

- *Degree of Naturalness*. The structure of a scene strongly differs between man-made and natural environments. Straight horizontal and vertical lines dominate man-made structures whereas most natural landscapes have textured zones and undulating contours. Therefore, scenes having a distribution of edges commonly found in natural landscapes would have a *high degree of naturalness* whereas scenes with edges biased toward vertical and horizontal orientations would have a *low degree of naturalness*.
- *Degree of Openness*. A second major attribute of the scene spatial envelope is its sense of Enclosure.[2] A scene can have a closed spatial envelope full of visual references (e.g., a forest, a mountain, a city center), or it can be vast and open to infinity (e.g., a coast, a highway). The existence of a horizon line and the lack of visual references confer to the scene a high *degree of Openness*. *Degree of Openness* of a scene decreases when the number of boundary elements increases.
- *Degree of Roughness*. Roughness of a scene refers principally to the size of its major components. It depends upon the size of elements at each spatial scale, their abilities to build complex elements and their relations between elements that are also assembled to build other structures, and so on. Roughness is correlated with the fractal dimension of the scene and thus, its complexity.
- *Degree of Expansion*. Man-made structures are mainly composed of vertical and horizontal structures. However, according to the observer's point of view, structures can be seen under different perspectives. The convergence of parallel lines gives the perception of the depth gradient of the space. A flat view of a building would have a *low degree of Expansion*. On the contrary, a street with long vanishing lines would have a *high degree of Expansion*.
- *Degree of Ruggedness*. *Ruggedness* refers to the deviation of the ground with respect to the horizon (e.g., from open environments with a flat horizontal ground level to mountainous landscapes with a rugged ground). A rugged environment produces oblique contours in the picture and hides the horizon line. Most of the man-made environments are built on a flat ground. Therefore, rugged environments are mostly natural.

Computations of *degree of openness* and *degree of roughness* would apply to any type of real world scenes. However, *degree of expansion* characterizes urban scenes better than natural landscapes. Convergence of parallel lines exists in natural landscapes but they are rare because of the lack of straight, long thin lines (as a counter example, a canyon may exhibit long corridors comparable to a long street). Similarly, *degree of ruggedness* characterizes natural landscapes (e.g., peaks and mountains) better than man-made scenes (to the exception of specific constructions such as the pyramids or some modern buildings). Therefore, the purpose of the spatial envelope model is to show that modeling these five spatial properties is adequate to perform a high-level description of the scene. In that regard, the next section defines the level of scene description we attempt to achieve as well as the image-based representation relevant for that level of description.

IV. Modeling the Scene Structure

In this section, we begin with an overview of the computational systems dedicated to scene recognition, and we present the different levels of description used for representing a scene. Then, we introduce the basis of the scene representations based on Fourier Transform and Principal Components Analysis. We show how the second order statistics of real world images are strongly constrained by the categories to which they belong.

A. Levels of Description and Scene Models

Scene models and the related computational approaches depend on the task to be solved (e.g., 3D reconstruction, object recognition, scene categorization) and the level of description required. Inspired by the terminology introduced by Rosch and Mervis (1975) for object description, the description of an environmental scene can be done at three different levels: Subordinate level (e.g., cars and people in a street), basic-level (e.g., a street), and superordinate level (e.g., an urban environment). Although not much attention has been paid to these levels of representation in computational vision, in particular in scene recognition models, these three descriptions provide different levels of abstraction and thus, different semantic information. Consequently, these three levels of description require different computational approaches.

Subordinate Level. This level of description requires the analysis of local structures, the recognition of objects or the labeling of regions in the image (e.g., grass, sky, building, people). In the framework of image retrieval, Carson et al. (1997, 1999) have proposed an image representation, termed *Blobworld*, that recognizes the image as a combination of objects. Basically, after segmenting the image into regions well specified in terms of texture and color, the system searches for images in the database with similar configurations and sizes of the constituent regions. Consequently, the internal representation given by Blobworld is typically performed at the level of objects, so the system performs much better when searching for distinctive objects in a simple background (e.g., pictures with faces, a specific animal in its natural environment) than when searching for more abstract categories. De Bonet and Viola (1997) proposed a different approach for scene representation that could also refer to a subordinate description. The image is represented by a high dimensional features vector obtained from the output of a tree of non-linear filters. Their system retrieves specific pictures of objects based on the similarity between regions with particular textural, spatial and color properties (e.g., sport cars, sunsets). But their method, based as it is on a very high dimensional signature, does not allow the formation of an internal meaningful representation of the scene.

Basic Level. The basic level categorization corresponds to the most common categorical representation (e.g. forest, mountain, street). Members of a basic level category usually have a similar shape (e.g. similar components) and share the same function. In that regard, Lipson et al. (1997) encoded the global configuration of a scene by using spatial and photometric relationships within and across predetermined regions of images. They show that the design of flexible spatial templates can successfully classify natural scenes with the constraint that the categories are geometrically well defined (e.g., blue blob above white blob above brown blob, for a snowy mountain).

Superordinate level. This level of description corresponds to the highest level of abstraction, and therefore, it has the lowest visual category resemblance. Several studies have focused on this level of description: Gorkani and Picard (1994) classified pictures into two categories (cities and natural landscapes) based on the statistics of orientations in the image. Szummer

and Picard (1998) discriminated between indoor and outdoor environments based on color and texture features. Vailaya et al. (1998, 1999) have proposed a more complete set of categories (indoor, outdoor, city, landscapes, and sunsets, mountains and forests). Common to all of these methods is the use of a scene representation based on color and edge statistics that classify scenes in exclusive classes.

In this paper, we intend to represent the structure of a scene image at both a superordinate level and a basic level. We will first use the spatial envelope attributes for building an abstract description of the scene (e.g., natural, open, expanded, small, among others) and we will show that the spatial envelope attributes provide a meaningful description of the space that the image subtends (e.g., perspective view of a large urban space with small elements) that allows inference of its probable basic level category (e.g., street). For now, we present in the following section, the image-based representation.

B. Image-Based Representations

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

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

$i(x, y)$ is the intensity distribution of the image[3] along the spatial variables (x, y), f_x and f_y are the spatial frequency variables. $h(x, y)$ is a circular Hanning window to reduce boundary effects. Due to the spatial sampling, $I(f_x, f_y)$ is a periodic function. The central period is $(f_x, f_y) \in [-0.5, 0.5] \times [-0.5, 0.5]$, units are in cycles per pixel. The complex function $I(f_x, f_y)$ is the Fourier transform that can be decomposed into two real terms: $A(f_x, f_y) = |I(f_x, f_y)|$, the amplitude spectrum of the image, and $\Phi(f_x, f_y)$, the phase function of the Fourier transform.

The phase function $\Phi(f_x, f_y)$ represents the information relative to the local properties of the image. It contains information relative to the form and the position of image components (Morgan et al., 1991, Piotrowski and Campbell, 1982). By contrast, $A(f_x, f_y)$ gives unlocalized information about the image structure: the amplitude spectrum represents the spatial frequencies spread everywhere in the image, and thus informs about the orientation, smoothness, length and width of the contours that compose the scene picture. The squared magnitude of the Fourier transform (energy spectrum) gives the distribution of the signal's energy among the different spatial frequencies. Therefore, the energy spectrum provides a scene representation invariant with respect to object arrangements and object identities, encoding only the dominant structural patterns present in the image. Previous studies have shown that such unlocalized information can be relevant for simple classification tasks (e.g., Gorkani and Picard, 1994; Guerin and Oliva, 2000; Oliva et al., 1999; Torralba and Oliva, 1999, submitted; Szummer and Picard, 1998; Vailaya et al., 1998, 1999). In the next section, we will provide more evidence that the statistics of unlocalized spectral features are strongly constrained for several scene categories.

Another relevant piece of information for image representation concerns the spatial relationships between the main structures in the image (e.g., Carson et al., 1997, 1999; De Bonet and Viola, 1997; Lipson et al., 1997; Torralba and Oliva, 1999). Spatial distribution of spectral information can be described by means of the windowed Fourier transform (WFT):

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

where $h_r(x', y')$ is a hamming window with a circular support of radius r. The localized energy spectrum (spectrogram), $A(x, y, f_x, f_y)^2 = |I(x, y, f_x, f_y)|^2$, provides localized structural information and it can be used for a detailed analysis of the scene by using a small size window. As one of the goals of this study is to show that scene categorization can be achieved bypassing object recognition stages, we chose a representation with a poor spatial resolution (e.g., Carson et al., 1999; Lipson et al., 1997; Torralba and Oliva, 1999). More specifically, we computed the WFT at 8×8 spatial locations with large overlapping neighborhoods, with a diameter of 64 pixels each.

Both the global energy spectrum and the spectrogram provide high dimensional representations of the input image $i(x, y)$. Common techniques used in pattern recognition for feature extraction and dimensionality reduction are the Karhunen-Loeve Transform (KLT) and the Principal Component Analysis (PCA). The KLT yields a decomposition of a random signal by a set of orthogonal functions with decorrelated coefficients. Dimensionality reduction is achieved by the PCA by

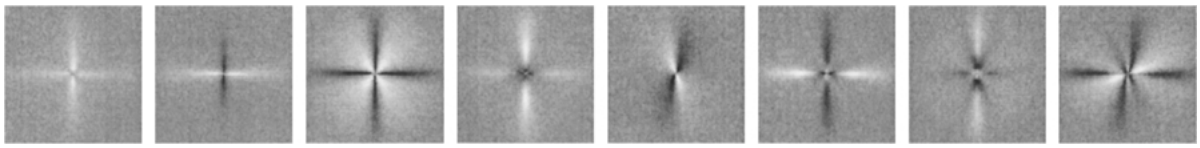

Figure 2. The first eight principal components for energy spectra of real-world scenes. The frequency $f_x = f_y = 0$ is located at the center of each image.

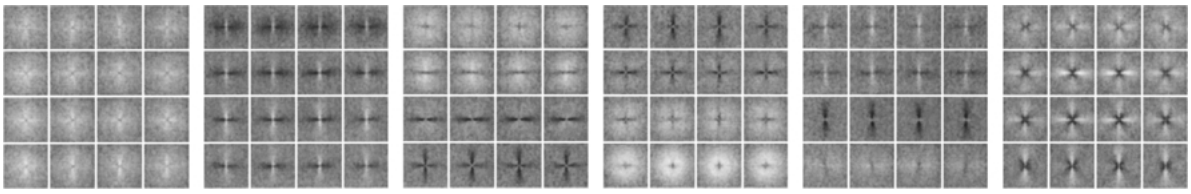

Figure 3. The first six principal components of the spectrogram of real-world scenes. The spectrogram is sampled at 4×4 spatial location for a better visualization. Each subimage corresponds to the local energy spectrum at the corresponding spatial location.

considering only the KL functions that account for the maximal variability of the signal described.

The energy spectrum can be decomposed into a KL basis as follows:

$$A(f_x, f_y)^2 \simeq \sum_{i=1}^{N_G} v_i \, \psi_i(f_x, f_y) \tag{3}$$

and, similarly, the spectrogram:

$$A(x, y, f_x, f_y)^2 \simeq \sum_{i=1}^{N_L} w_i \, \Psi_i(x, y, f_x, f_y) \tag{4}$$

N_G and N_L are the number of functions used for the approximations and determine the dimensionality of each representation. The coefficients v_i and w_i of the decompositions are obtained as:

$$\begin{aligned} v_i &= \langle A^2, \psi_i \rangle \\ &= \iint A(f_x, f_y)^2 \, \psi_i(f_x, f_y) \, df_x \, df_y \end{aligned} \tag{5}$$

and

$$\begin{aligned} w_i = \sum_x \sum_y \iint & A(x, y, f_x, f_y)^2 \\ & \times \Psi_i(x, y, f_x, f_y) \, df_x \, df_y \end{aligned} \tag{6}$$

$\psi_i(f_x, f_y)$ are the KL basis of the energy spectrum and verify orthogonality $\langle \psi_i, \psi_j \rangle = 0$ and decorrelation of the decomposition coefficients $E[v_i \, v_j] = 0$ for $i \neq j$. The functions $\Psi_i(x, y, f_x, f_y)$ are the KL basis of the localized energy spectrum providing also an orthogonal basis with decorrelated decomposition coefficients. Figures 2 and 3 show the KL basis obtained (see appendix). The visualization provides a simple interpretation of how the coefficients v_i and w_i are obtained from the energy spectrum and the spectrogram respectively. The KL decompositions of the energy spectrum and the localized energy spectrum provide two sets of features for representing the scene structure:

- $\mathbf{v} = \{v_i\}_{i=1,N_G}$: provides unlocalized structural information. $\mathbf{v}$ contains a low-resolution description of the energy spectrum of the image.
- $\mathbf{w} = \{w_i\}_{i=1,N_L}$: provides structural information with a description of the spatial arrangement. Due to the reduced dimensionality ($N_L < 50$), the representation has a low resolution in both spectral and spatial domains.

The two representations are redundant as the information provided by $\mathbf{v}$ is contained in the more complete description provided by $\mathbf{w}$. Therefore, only the representation performed by the WFT is required. Nevertheless, the unlocalized spectral information provides a simple way to represent the frequency components that dominate the whole image while it may help to understand the basic structural differences between environments of different sorts. It must be noted that both representations $\mathbf{v}$ and $\mathbf{w}$ are holistic as they encode the whole image without splitting it into objects or regions.

We will provide through the paper, results for both the unlocalized (energy spectra) and localized (spectrogram) representations with the aim to measure

their respective contribution to the scene representation. In Sections V and VI, we will show that even unlocalized structural information is capable of providing reliable information about the spatial envelope properties of a scene and its category, although more accurate results are obtained from the WFT. But, for now, in order to illustrate the nature of the structural information that differentiates scene categories, we review in the next section studies in the field of image statistics and in particular, with a focus into the second order statistics (energy spectrum) of real-world images. We have extended those studies showing that real-world images corresponding to different categories have very different second order statistics, and thus global structure. Therefore, we can expect that the representations **v** and **w** may provide discriminant information about the spatial envelope and the scene category.

C. *Spectral Signature of Scene Categories*

Studies devoted to the statistics of real-world images have observed that the energy spectra of real-world images fall in average with a form $1/f^{\alpha}$ with $\alpha \sim 2$ (or $\alpha \sim 1$ considering the amplitude spectrum). The average of the energy spectrum provides a description of the correlation found in natural images (Field, 1987, 1994; van der Schaaf and van Hateren, 1996), and it has several implications for explaining the processing carried out by the first stages of the visual system (Field, 1987; Atick and Redlich, 1992). In that regard, a few studies have shown that different kinds of environments exhibit very specific and distinctive power spectrum forms (e.g., Baddeley, 1997; Oliva et al., 1999; Switkes et al., 1978).

In order to illustrate the structural aspects that are captured by the energy spectrum, we computed the spectral signatures of the following basic level scene categories: tall buildings, highways, city close-up views and city centers for man-made environments, and coasts, mountains, forests and close-up views for natural scenes. The spectral signatures were computed by averaging the energy spectrum of hundreds of exemplars for each category. The spectral signatures can be adequately approximated by a function:

$$E[A(f,\theta)^2 \mid S] \simeq \Gamma_s(\theta)/f^{-\alpha_s(\theta)} \qquad (7)$$

where $E[A(f,\theta)^2 \mid S]$ is the expected value of the energy spectrum for a set of pictures belonging to the category S. Spatial frequencies are represented in polar coordinates (f,θ). Functions $\Gamma(\theta)$ and $\alpha(\theta)$ are obtained by a linear fitting of the averaged energy spectrum on logarithmic units for each orientation θ (see van Der Schaaf and van Hateren (1996) for a detailed analysis). Figure 4 shows examples of the linear fitting for different orientations for three scene categories and Fig. 5 shows the spectral signatures of the eight scene categories. The model of Eq. (7) provides correct fitting for all the eight categories for frequencies below 0.35 cycles/pixel (as noise and aliasing corrupt higher spatial frequencies, see Fig. 4).

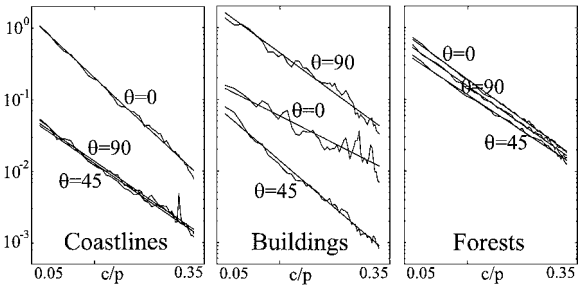

Figure 4. Examples of sections at different orientations of the averaged energy spectrum for three scene categories, and the corresponding linear fitting.

The functions $\Gamma(\theta)$ and the function $\alpha(\theta)$ are related to different perceptual features. The function $\Gamma(\theta)$ reveals the dominant orientations of a scene category (see Fig. 5). The function $\alpha(\theta)$, represents the slope of the decreasing energy spectrum values, from low to high spatial frequencies. The slope varies as a function of the complexity of the scene. Pentland (1984) showed that fractal natural surfaces (as mountains, forests) produce a Fractal image with an energy spectrum of the form $1/f^{\alpha}$, where α is related to the fractal dimension of the 3D surface (e.g., its roughness). Slope characteristics may be grouped in two main families: a slow slope ($\alpha \sim 1$) for environments with textured and detailed objects and a steep slope ($\alpha \sim 3$) for scenes with large objects and smooth edges. The slower is the slope, the more textured the image is. Examples of scenes categories with different slopes and therein with different roughness, are shown in Fig. 5(c) and (d) and (g) and (h). Even thought they have similar dominant orientations $\Gamma(\theta)$, their spectral signatures differ in the function $\alpha(\theta)$.

When considering a large number of real-world scenes without differentiating among different categories, the images have stationary statistics. However, in contrast to images of textures where most of the statistics are stationary regardless of the category,

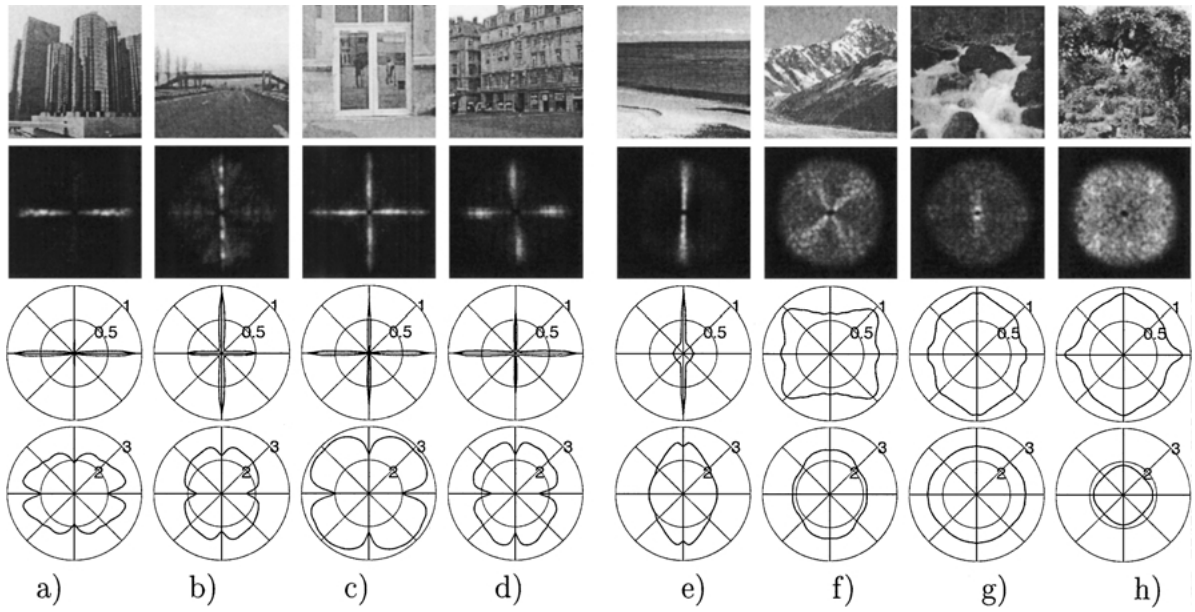

Figure 5. Examples of scenes from different categories, their respective energy spectrum (energy spectra have been multiplied by f^2 in order to enhance the visibility of high spatial frequencies) and the spectral signatures of their category: function $\Gamma_s(\theta)$ and the bottom line shows the function $\alpha_s(\theta)$ in a polar diagram. From a) to h), scenes illustrate the categories: tall building, highway, urban close-up views, city center, coast, mountain, natural close up views and forests.

environmental scenes belonging to the same category are characterized by particular arrangements of structures within the image (Lipson et al., 1997; Torralba and Oliva, 1999). For instance, a street is composed of the road, buildings and the sky that are arranged in a very predictive way. This arrangement of image regions with different structural characteristics that is typical of the category *street* introduces a spatial non-stationary behavior of the statistics of the image when considering a large set of images belonging to the same scene category. This non-stationary behavior is typical of several scene categories and provides relevant information for the determination of the category of which a scene picture belongs. The spatial non-stationary behavior of the second order statistics can be studied by the spectrogram as introduced in Section IV(b). Figure 6 shows the mean spectrogram obtained from averaging the spectrogram of hundreds of scene pictures belonging to the same category. The categories shown in Fig. 6 are: man-made open (a) and urban vertically structured (b) environments, perspective views of streets (c), far view of city-center buildings (d) and close-up views of outdoor urban structures (e) and natural open (f) and enclosed (g) environments, mountainous landscapes (h), enclosed forests (i) and close-up views of non-textured natural structures like rocks and water (j). It must be noted that non-stationarity is a characteristic of open environments (a) and (f) and semi open environments as (b), (c) and (h). Open and semi open environments, which correspond to large spaces, have strong organization rules of their main scene components (support surfaces, horizon line, sky, vertical and textured structures). However, enclosed environments (d), (e), (g), (i) and (j) are almost stationary in the second order statistics. Enclosed environments (as forests, small urban spaces, etc.), although they differ in the constituent structural elements (energy spectrum), do not have very strong organizational rules.

V. Estimation of the Spatial Envelope Properties

The structural differences between scene categories provide a cue for scene recognition that does not require previous region labeling or individual object recognition. The goal of this section is the estimation of the spatial envelope attributes from the two spectral representations. In particular, we will look for the spectral attributes and the spatial organizations that are correlated with the spatial envelope attributes.

The principal components of each of the two image representations (the global energy spectrum and the

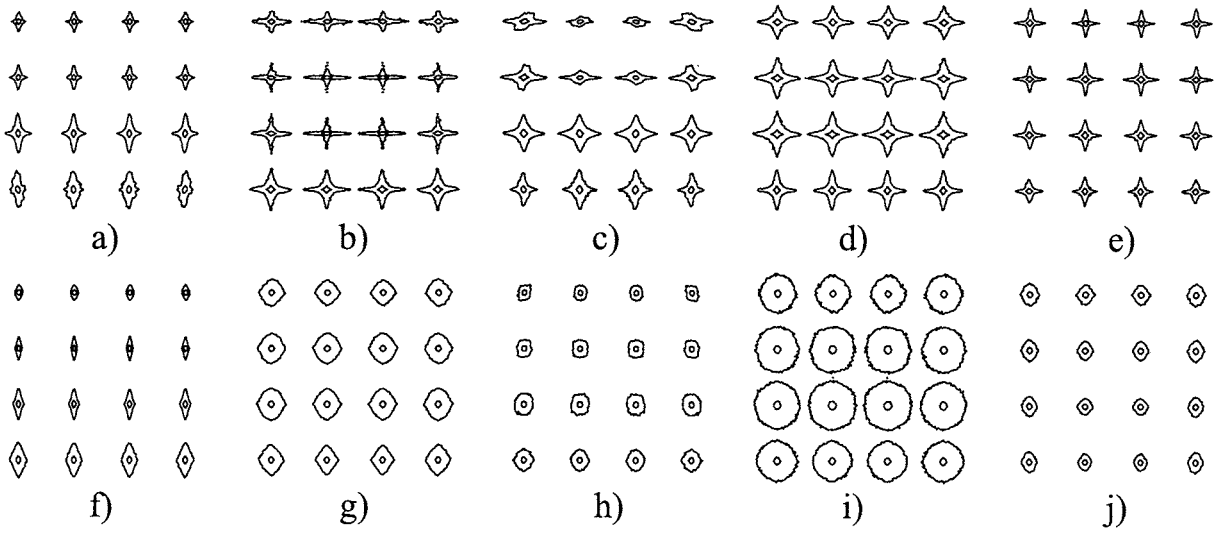

Figure 6. Examples of sections of the mean spectrogram (see the text for a detailed description).

spectrogram), define an image-based feature space into which each scene can be projected. However, the contribution of each feature cannot be understood as they stand, and more importantly, they are not directly meaningful to human observers. Spatial envelope properties represent the scene in a very low dimensional space in which each dimension depicts a meaningful property of the space of the scene. Therein, it is possible to assign a specific interpretation to each dimension: along the openness dimension, the image refers to an open or a closed environment; along the roughness dimension, the scene refers to an environment made with small vs. large elements (this relates to the complexity of the scene and the size of the space), etc.

A. *Discriminant Spectral Templates*

The estimation of the spatial envelope attributes from image-based features can be solved using different regression techniques. In the case of a simple linear regression, the estimation of a scene attribute s from the global spectral features $\mathbf{v}$ of a scene picture can be written as:

$$\hat{s} = \mathbf{v}^T \mathbf{d} = \sum^{N_G} v_i \, d_i = \iint A(f_x, f_y)^2 \, DST(f_x, f_y) \, df_x \, df_y \tag{8}$$

with (from Eq. (6)):

$$DST(f_x, f_y) = \sum^{N_G} d_i \, \psi_i(f_x, f_y) \tag{9}$$

Although more complex non-linear regression models can be used (e.g., mixtures of experts, etc. See Ripley, 1996), the linearity of the operations in Eq. (8) provides a simple writing of the estimation process giving a simple interpretation of the system behavior. For instance, Eq. (8) shows that the spatial envelope property s is estimated by a dot product between the amplitude spectrum of the image and a template $DST(f_x, f_y)$. The DST (Discriminant Spectral Template) is a function that describes how each spectral component contributes to a spatial envelope property (Oliva et al., 1999; Torralba and Oliva, 1999). The DST is parameterized by the column vector $\mathbf{d} = \{d_i\}$ which is determined during a learning stage detailed below.

A similar estimation can be performed when using the spectrogram features $\mathbf{w}$:

$$\hat{s} = \mathbf{w}^T \mathbf{d} = \sum_{i=1}^{N_L} w_i \, d_i = \sum_x \sum_y \iint A(x, y, f_x, f_y)^2 \times WDST(x, y, f_x, f_y) \, df_x \, df_y \tag{10}$$

with:

$$WDST(x, y, f_x, f_y) = \sum_{i=1}^{N_L} d_i \, \Psi_i(x, y, f_x, f_y) \tag{11}$$

The *WDST* (Windowed Discriminant Spectral Template) describes how the spectral components at different spatial locations contribute to a spatial envelope property (Torralba and Oliva, 1999). The sign of the values of the *WDST* indicates the sign of the correlation between the spectral components and the spatial envelope property s.

In order to determine the parameters $\mathbf{d}$ related to a specific property of the spatial envelope, we used the following learning procedure: we selected a random set of 500 scene pictures from the database (see section V.B) and we placed them along an axis sorted according to the spatial envelope property that we wanted to estimate. The training set consists in the features vectors $\{\mathbf{v}_t\}_{t=1,500}$ (or $\{\mathbf{w}_t\}_{t=1,500}$) and the corresponding values of the spatial envelope property $\{s_t\}_{t=1,500}$ given by the location of each picture along the axis. For each picture, we estimate the attribute as $\hat{s}_t = \mathbf{v}_t^T\mathbf{d} + d_0$, where d_0 is a constant. The constant d_0 is not considered in Eqs. (8) and (10) as it does not affect the organization and the discrimination performances of the attribute. The parameters vector $\mathbf{d}$ that minimize the mean squared error is (e.g., Ripley, 1996):

$$\mathbf{d}_1 = \left(\mathbf{V}_1\mathbf{V}_1^T\right)^{-1}\mathbf{V}_1\mathbf{s} \tag{12}$$

The column t of the matrix $\mathbf{V}_1$ corresponds to a vector composed by the features vector of the image t and a 1: $[\mathbf{v}_t;\ 1]$. The vector $\mathbf{d}_1$ contains the DST parameters (or WDST) and the constant term $d_0 : \mathbf{d}_1 = [\mathbf{d};\ d_0]$. The inversion of the matrix $(\mathbf{V}_1\mathbf{V}_1^T)^{-1}$ may be ill conditioned if the number of spectral features used for the learning (N_G or N_L) is too large.

The regression procedure is appropriate for attributes that organize scenes in a continuous manner (e.g., degree of openness, expansion, ruggedness, and roughness). However, some scene properties refer to a binary classification (e.g., man-made vs. natural, indoor vs. outdoor, objects vs. environments, etc.). The discrimination of two classes can be performed by assigning to the images of each class the attribute values $s_t = -1$ or $s_t = 1$ for the two classes respectively. In such a case, the regression parameters (Eq. (12)) are equivalent to the parameters obtained by applying a linear discriminant analysis (see Ripley, 1996; Swets and Weng, 1996).

All the computations presented are performed in the frequency domain. However, it would be interesting to localize in the image itself the spatial features that contribute to the estimation of each spatial envelope property s. Equation (8) shows how the attribute is computed from the energy spectrum. As the $DST(f_x, f_y)$ function contains both positive and negative values, we first separate the *DST* into two positive functions: $DST = DST_+ - DST_-$, with $DST_+(f_x, f_y) = rect[DST(f_x, f_y)]$ and $DST_- = rect[-DST(f_x, f_y)]$. $rect(x) = x$ for $x > 0$ and $rect(x) = 0$ for $x \leq 0$. From Eq. (8) we obtain:

$$\hat{s} = \iint A(f_x, f_y)^2\, DST_+(f_x, f_y)\, df_x\, df_y - \iint A(f_x, f_y)^2\, DST_-(f_x, f_y)\, df_x\, df_y \tag{13}$$

This equation interprets s as computed by the difference between the output energies of two filters with transfer functions: $|H_+(f_x, f_y)|^2 = DST_+(f_x, f_y)$ and $|H_-(f_x, f_y)|^2 = DST_-(f_x, f_y)$. Parseval equality applies for each integral (energy can be computed in the spatial domain):

$$\hat{s} = \sum_{x,y}[i(x, y) * h_+(x, y)]^2 - \sum_{x,y}[i(x, y) * h_-(x, y)]^2 \tag{14}$$

$h_+(x, y)$ and $h_-(x, y)$ are the impulse responses of two filters with the transfer functions: $H_+(f_x, f_y)$ and $H_-(f_x, f_y)$. $i(x, y)$ is the input image and $*$ is the convolution operator. The functions h_+ and h_- are not uniquely constrained by the *DST* as the phase function can have any value. We fix the phase function at zero in order to have localized spatial functions. The image composed by

$$a(x, y) = [i(x, y) * h_+(x, y)]^2 - [i(x, y) * h_-(x, y)]^2 \tag{15}$$

shows how each spatial location contributes to the attribute $s : \hat{s} = \sum_{x,y} a(x, y)$. We refer to $a(x, y)$ as an *opponent energy image*. The functions $h_+(x, y)$ and $h_-(x, y)$ give the shape of the most discriminant spatial features used for computing s.

In a similar vein, we can derive the contribution of each spatial location for the estimation of the attribute s using the spectrogram. In this case, we obtain two spatially variant filters $h^+_{x',y'}(x, y)$ and $h^-_{x',y'}(x, y)$ with spatially variant transfer functions $|H^+_{x',y'}(f_x, f_y)|^2 = WDST_2 + (x', y', f_x, f_y)$ and $|H^-_{x',y'}(f_x, f_y)|^2 = WDST_-(x', y', f_x, f_y)$. The variables (x', y')

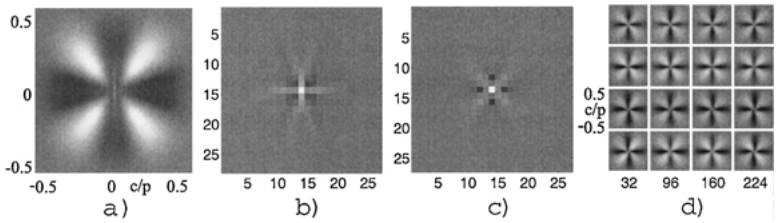

Figure 7. Discriminant spectral templates for the degree of naturalness: a) $DST(f_x, f_y)$, units are in cycles per pixel, b) $h_-(x, y)$, units are in pixels, c) $h_+(x, y)$ and d) $WDST(x, y, f_x, f_y)$ sampled at 4×4 spatial locations. White and dark pixels correspond, respectively, to positive and negative values.

refer to the spatial locations at which the spectrogram has been sampled. The *opponent energy image* is then computed as:

$$a(x, y) = \sum_{x'} \sum_{y'} ([i_{x',y'}(x, y) * h^{+}_{x',y'}(x, y)]^2 - [i_{x',y'}(x, y) * h^{-}_{x',y'}(x, y)]^2) \quad (16)$$

with $i_{x',y'}(x, y) = i(x, y)h_r(x - x', y - y')$. The variables x' and y' correspond to the spatial locations at which the spectrogram has been sampled (8×8 spatial locations for the rest of the paper).

To summarize, the definition of the spatial envelope properties provides a low dimensional representation (much lower dimensionality than image-based features such as wavelet descriptors). This method proposes a reduced number of filters/wavelets relevant for the scene recognition task. The next sections are devoted to the computation of the spatial envelope properties.

B. Environmental Scene Database

The database contains about 8100 pictures of environmental scenes so as to cover a large variety of outdoor places. Images were 256×256 pixels in size, in 256 gray levels. They come from the Corel stock photo library, pictures taken from a digital camera and images downloaded from the web. The scene database was composed of about 4000 natural scenes (e.g., coast, beach, ocean, island, field, desert, grassland, valley, lake, river, mountains, canyon, cavern, forest, waterfall, garden, etc.), and about 3500 urban environments (e.g., skyscraper, city center, commercial area, street, road, highway, house, building, pedestrian center, place, parking, etc.). The rest of images ($\simeq$600) correspond to ambiguous scenes in terms of degree of naturalness (e.g., farming scene, village in mountains, panoramic and aerial city views, etc.).

C. Degree of Naturalness

2000 images of natural and man-made scenes[4] were used for computing the naturalness DST and WDST according to the learning procedure described Section V(a). Ambiguous scenes in terms of naturalness (images with both man-made and natural structures) were not used in the learning. As the naturalness decision is almost a binary categorization, the linear discriminant analysis instead of the regression has been used for the learning stage. The resulting $DST(f_x, f_y)$ and $WDST(x, y, f_x, f_y)$ are presented in Fig. 7(a) and (d). The $DST(f_x, f_y)$ shows how the spectral components of each scene energy spectrum should be weighted in order to discriminate whether the image is a natural or a man-made environment. As shown in Section IV, man-made scenes exhibit a higher proportion of H and V orientations (see Fig. 5(a)–(d)). The dark negative parts show that this predominance arises mostly at medium and high spatial frequencies. The white parts are associated with a *high degree of naturalness* and represent low spatial vertical contours and diagonals at almost all the spatial scales. Figure 7 also shows the two filters $h_-(x, y)$ and $h_+(x, y)$ that are the spatial equivalents to the naturalness DST. $h_-(x, y)$ cancels oblique orientations and enhances cross junctions aligned with the horizontal and vertical directions of man-made scenes. $h_+(x, y)$ is matched to oblique orientations and cancels horizontal and vertical edges.[5]

To test the validity of the templates, about 5000 scenes not used in the learning stage were projected onto the $DST(f_x, f_y)$ and the $WDST(x, y, f_x, f_y)$. On average, 93.5% of man-made scenes and natural landscapes were correctly classified, for both procedures (see a sample of misclassified scenes, Fig. 8). Performances do not differ when using a Bayesian classifier (mixture of gaussians, e.g., Ripley, 1996). To illustrate the classification performances, Fig. 9 shows a sample of scenes selected at random, and then projected onto

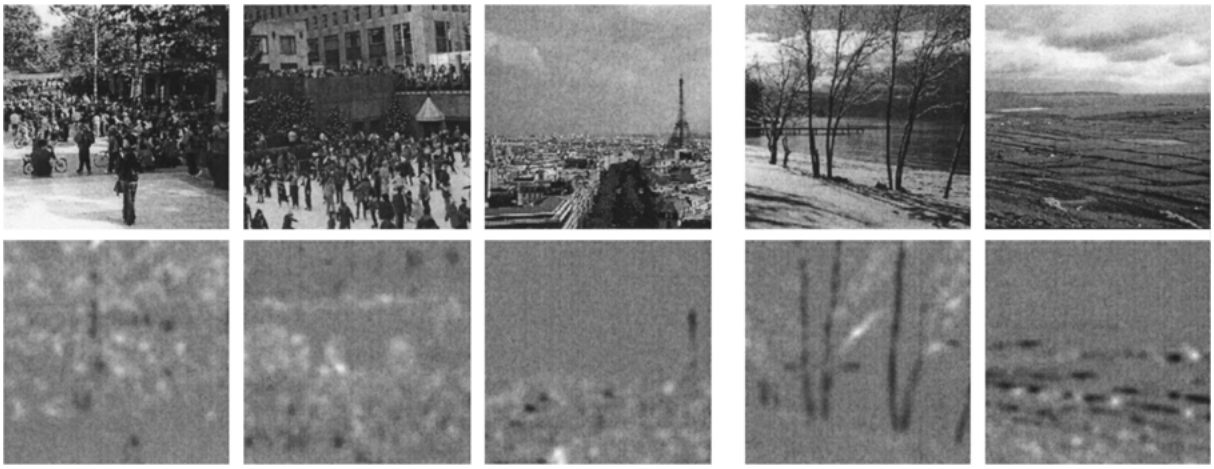

Figure 8. Samples of scenes (top) not correctly classified with the DST, and their opponent energy image (bottom). Errors of man-made scenes mostly include textured scenes. Errors of natural landscapes mostly include forests with vertical trees and some open landscapes having straight lines.

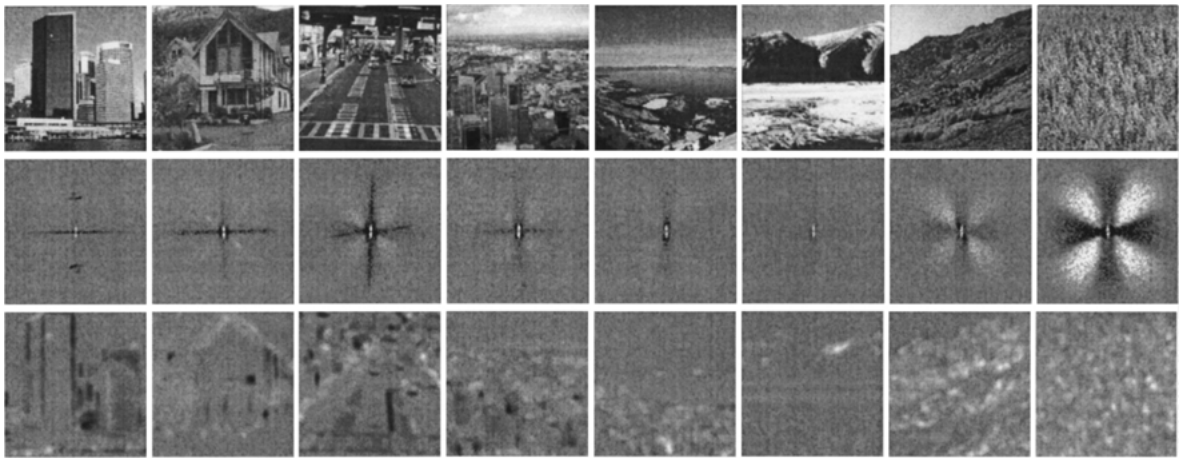

Figure 9. From top to bottom: Samples of images selected at random ordered along the naturalness axis, from man-made environments (left) to natural landscapes (right); their energy spectra multiplied by the DST; the opponent energy image (we have suppressed the effect of the Hanning window for clarity). Natural and man-made components are respectively represented by white and black edges.

the DST (very similar organizations are observed when using the WDST instead). The middle line of Fig. 9 shows the product $A(f_x, f_y)^2\, DST(f_x, f_y)$, that illustrates how the energy spectrum is weighted for each image. The bottom line of Fig. 9 shows the functions $a(x, y)$, or opponent energy images. The organization of images along the naturalness axis evolves according to what the discriminant analysis has considered as relevant for separating at best the two groups. The spectral evolution goes from scenes with the *lower degree of naturalness* (e.g., scenes with a straight horizontal energy spectrum form followed by cross-like energy spectrum forms), to scenes with the *higher degree of naturalness* (e.g., scenes with a vertical energy spectrum followed by isotropic energy spectrum forms). The naturalness DST and WDST represent to which degree a scene picture is closed to a natural (vs. man-made) environment. This representation does not mean that a skyscraper view is a better exemplar of the man-made category than a city center view but that the probability that a vertically structured scene represents a man-made scene is higher than a scene with horizontal contours or isotropic components.

Results of classification show how powerful the global energy spectrum (DST) may be for resolving the man-made vs. natural distinction. In fact, the

introduction of spatial information does not seem to improve the classification. Figure 7(d) shows that the resulting WDST does not vary with respect to the spatial location. As natural and man-made environmental categories cover (almost) all the possible spatial arrangements of their main components, the second order statistics are stationary. Therefore, the resulting WDST is also stationary. The naturalness WDST is simply a replication of the shape of the global naturalness DST at each location. But, as detailed hereafter, the stationary property of the WDST varies with the spatial envelope attribute estimated.

As explained in Sections III and IV, as the spectral components correlated with the spatial envelope properties differ between natural and man-made environments, we computed the other spatial envelope properties independently for each type of environment.

D. Computation of the Spatial Envelope Properties of Natural Scenes

As openness, ruggedness and roughness properties are continuous dimensions, we used the linear regression procedure for the learning stage.[6] More precisely, 500 natural scenes were randomly selected among the 4000 natural scenes and then organized along the three dimensions as follows. For the openness property, we arranged scenes from widely open environments (e.g., coastlines, open landscapes) to enclosed environments (e.g., mostly forests), with other semi-open landscapes (fields, valleys, mountains, etc.) ordered between these two extremes. For estimating the ruggedness property, the scenes were arranged from natural scenes with long diagonals (mountains, valleys, peaks, etc.) to scenes complementary in terms of orientations: open scenes (e.g., coast, beach, field), vertical scenes (e.g., forest, falls) and scenes with isotropic texture (forest and textured landscapes). The roughness property corresponds to the global level of granularity of the scene surface that is correlated with the size of the elements or texture. We chose closed textured scenes for computing more accurately this dimension. Pictures were ordered from textured scenes made with small elements (mostly forests and fields) to coarse textured scenes (e.g., waterfalls, streams, rocks). The resulting DST and WDST are respectively shown in Figs. 10 and 11. The evaluation of each spectral template was assessed along two criteria: classification performances and ordering performances.

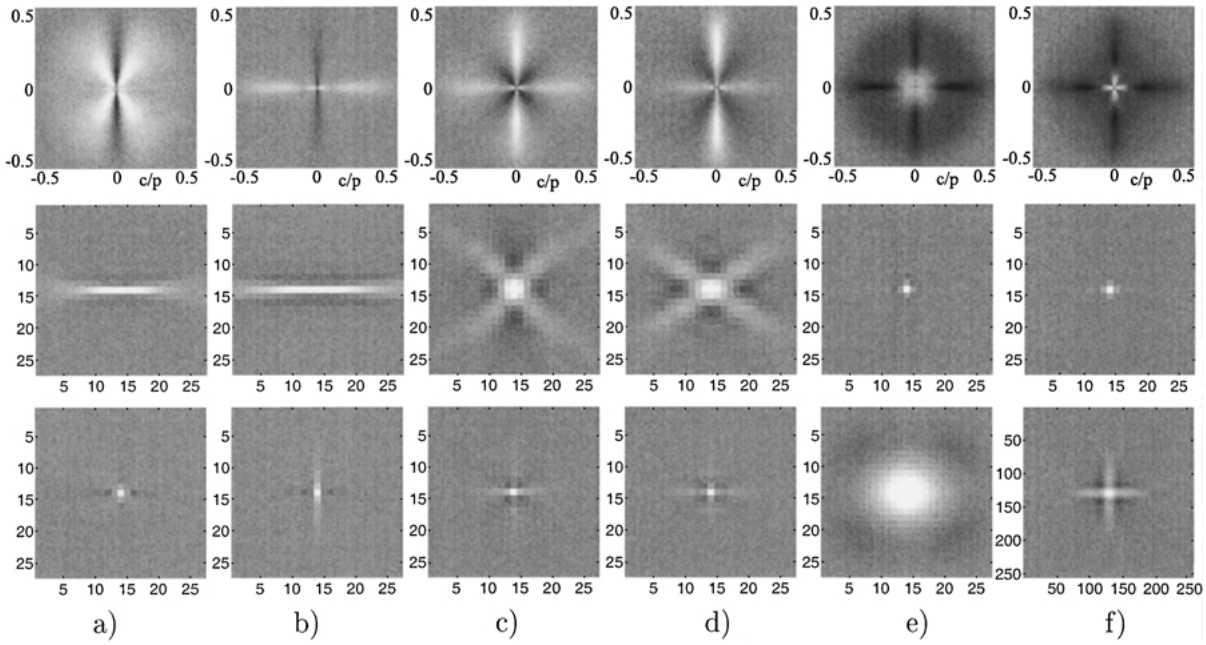

Figure 10. Discriminant spectral templates $DST(f_x, f_y)$, computed with $N_G = 16$, and the equivalent spatial feature detectors $h_-(x, y)$ and $h_+(x, y)$ for each property of the spatial envelope. Degree of openness for natural (a) and man-made scenes. (b) Degree of ruggedness for natural scenes (c). Degree of expansion for man-made scenes (d). Degree of roughness for natural (e) and man-made scenes (f).

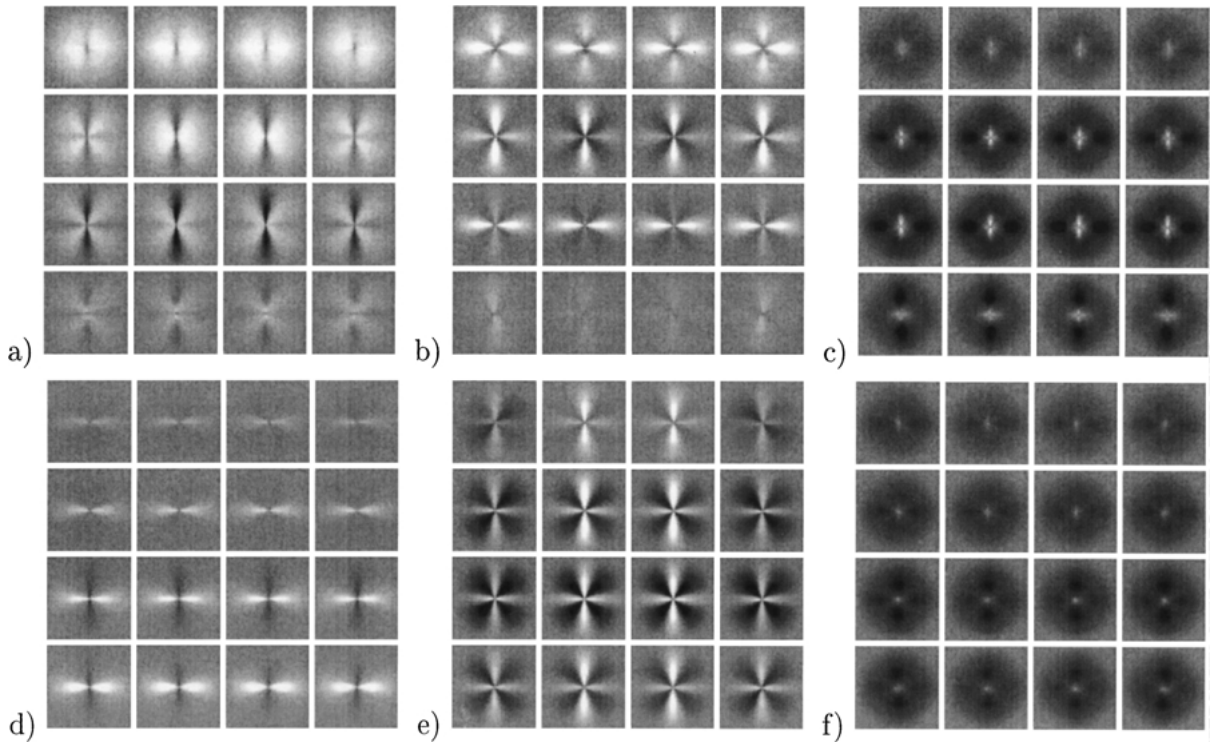

Figure 11. Discriminant spectral templates $WDST(x, y, f_x, f_y)$ with $N_L = 30$. For natural scenes: a) openness, b) ruggedness and c) roughness. For man-made scenes: d) openness, e) expansion and f) roughness.

The first criterion concerned the classification of natural scenes not used for the learning stage. These scenes were previously classified in one of the two exclusive groups per spatial envelope property (e.g., open vs. closed, fine vs. coarse texture). The openness DST (see Fig. 10(a)) clearly opposes vertical spectral components to other orientations. When projecting new open vs. closed scenes, accuracy was about 93%. In Fig. 10(a), we can see that the first filter (h_-) matches the existence of a horizon line in the scene and the second filter (h_+) corresponds to diagonals and isotropic distributions of orientations at medium and high spatial frequencies. The WDST (see Fig. 11(a)) shows how the spectral components of each scene energy spectrum should be locally weighted in order to organize scenes according to the openness property. As expected, the openness WDST clearly shows a different weighting at different spatial locations: the horizontal (dark vertical component) edges located around the center of the viewed scene are correlated with openness of the space (horizon line). By contrast, isotropic texture (white components) at the top of the scene indicates an enclosed space (forest, mountains). Categorization performances with the WDST were better (96.5%).

The ruggedness DST is displayed in Fig. 10(d). Basically, the DST opposes long diagonals to vertical and horizontal components at all the spatial frequencies. As shown by the WDST (Fig. 11(b)), the spatial arrangement of the spectral components seems to play a relevant role. Diagonal edges around the middle and top of the scene are correlated with a rugged environment (mountainous landscape). However, high spatial frequencies in the top and middle part and vertical edges are correlated with a flat ground level (e.g., forests and open environments). When projecting scenes (see examples in Fig. 12(b)) previously classified as having a high vs. low degree of ruggedness, accuracy was 89% with the DST and 91% with the WDST.

The roughness DST (Fig. 10(e)) is concerned with different slopes in the energy spectra. Scenes with a fine texture would match the dark zone of the DST, corresponding to high spatial frequencies (low slope value, $\alpha \sim 1.8$). Coarse textured scenes (high slope value, $\alpha \sim 2.5$) would match the low spatial frequencies (in white). The correct classification of new scenes belonging to fine vs. coarse textured scenes (see examples in Fig. 12(c)) was 92% with the DST and 94% with the WDST (Fig. 11(c)). Interestingly, the WDST

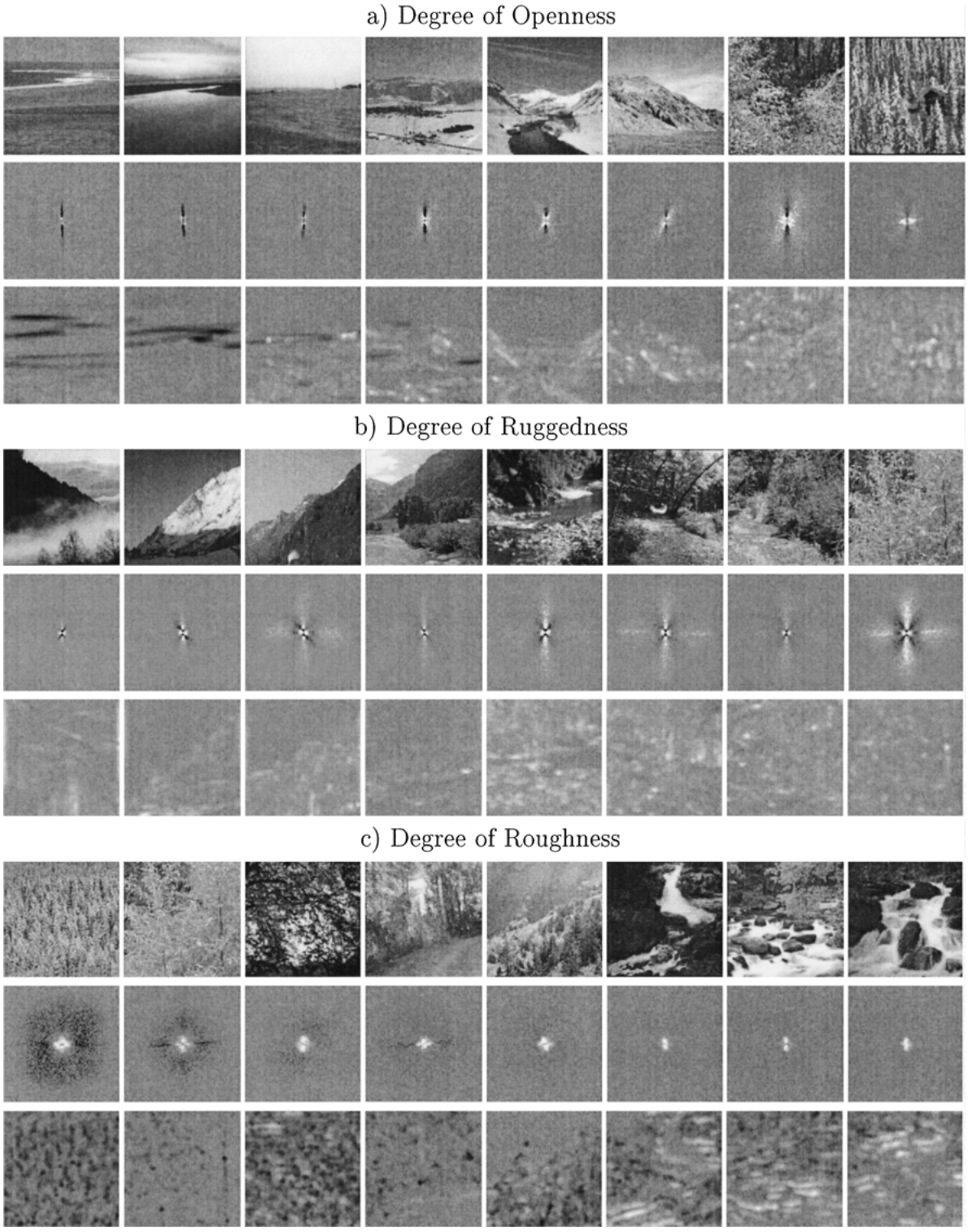

Figure 12. Samples of natural images selected at random and ordered with respect to their the degree of openness, degree of ruggedness, and degree of roughness. Each figure shows also the product $DST(f_x, f_y)A^2(f_x, f_y)$ and the opponent energy images $a(x, y)$ revealing the contribution of each spectral and spatial component to the computation of each attribute.

is stationary, weighting all the local spectral components in a similar way.

The second criterion we used for estimating the relevance of the template procedure concerned the organization of images along each spatial envelope dimension. If the templates are estimating the right weighting of the spectral features, scenes should be meaningfully arranged along an axis, according to the values of the estimated attribute of the spatial envelope. The ordering performed by each template was compared to human ordering. We asked four observers to perform 20 orderings, of 12 images each, for each of the three spatial envelope properties. Namely, subjects were told to begin the ordering by selecting the two pictures that were the most different according to one spatial envelope attribute and then to process by closed similarities from the two extreme. Orderings were compared by measuring the Spearman rank correlation:

$$Sr = 1 - 6\frac{\sum_{i=1}^{n}(rx_i - ry_i)^2}{n(n^2-1)} \qquad (17)$$

with $n = 12$. rx_i and ry_i are respectively the rank positions of the image i given by the algorithm and by one subject. A complete agreement corresponds to $Sr = 1$. When both orderings are independent, $Sr = 0$. A negative value of Sr means that the ordering has been inverted. We also computed the *Agreement* value that corresponds to the mean Spearman rank correlation between orderings given by the different subjects. Agreement evaluates the difficulty of the ordering task and the concordance of the criteria used by the subjects. Results of Table 2 show high correlations for the three spatial envelope properties. The average correlation between the DST (unlocalized spectral components) and the subjects is 0.79, and it increases to 0.84 when using localized spectral components (WDST). Furthermore, the average agreement among subjects is of 0.87, which indicates that orderings of 12 images performed by the template procedure or a human observer are very similar. Examples of scenes ordered by the templates along the openness, ruggedness and roughness dimensions are shown in Fig. 12. To illustrate the features (both spectral and spatial) that contribute to the estimation of each spatial envelope attribute, we show the results obtained with the DST (similar orderings are observed when using the WDST). More specifically, the middle row pictures in Fig. 12 show the product between the DST and the energy spectrum indicating the actual contribution of each spectral component to the estimated attribute. The opponent energy images (bottom part of Fig. 12) illustrate how the different spatial features of an image have been weighted.

Table 2. Correlation between orderings of natural scenes made by observers and the two templates for each spatial envelope property.

	Openness	Ruggedness	Roughness
DST	$m = 0.82$	0.73	0.82
WDST	$m = 0.88$	0.79	0.86
Agreement	0.92	0.82	0.87

Agreement measures the concordance between subjects.

E. *Computation of the Spatial Envelope of Urban Landscapes*

500 natural scenes were randomly selected from the 3500 man-made scenes and then organized along each spatial envelope property. For the openness property, urban environments were arranged from horizontally structured scenes (e.g. open roads, highways) to closed city views (e.g. streets, centers) and then vertically structured scenes (tall buildings, skyscrapers). For computing the degree of expansion, images were arranged from scenes with a perspective view (roads, some highways, streets) to "flat" scenes (e.g. views in front of buildings, shops, houses, at different distances). To build the Roughness template, the images were organized from "textured" urban scenes, very detailed, such as busy pedestrian scenes, some buildings with fine and small contours, to urban close-up views exhibiting larger surfaces (see exemplars in Fig. 13(c)). The templates were computed with the linear regression analysis.

The openness DST displayed in Fig. 10(b) has a very simple structure: vertical spectral components (in black) correspond to open scenes, as opposed to horizontal spectral components (in white) corresponding to vertical contours in the image correlated with an enclosed space. According to the shape of the DST, cross form energy spectra match both dark and white components. As a result, the corresponding scenes are located around the center of the axis (see Fig. 13(a)) between truly horizontal scenes and vertically structured scenes. Classification rate after projecting hundreds of open vs. closed and vertical scenes not used for the learning stage, was 94% with the DST and 96% with the WDST. Figure 11(d) shows the openness WDST for man-made scenes. The WDST reveals the non-stationary

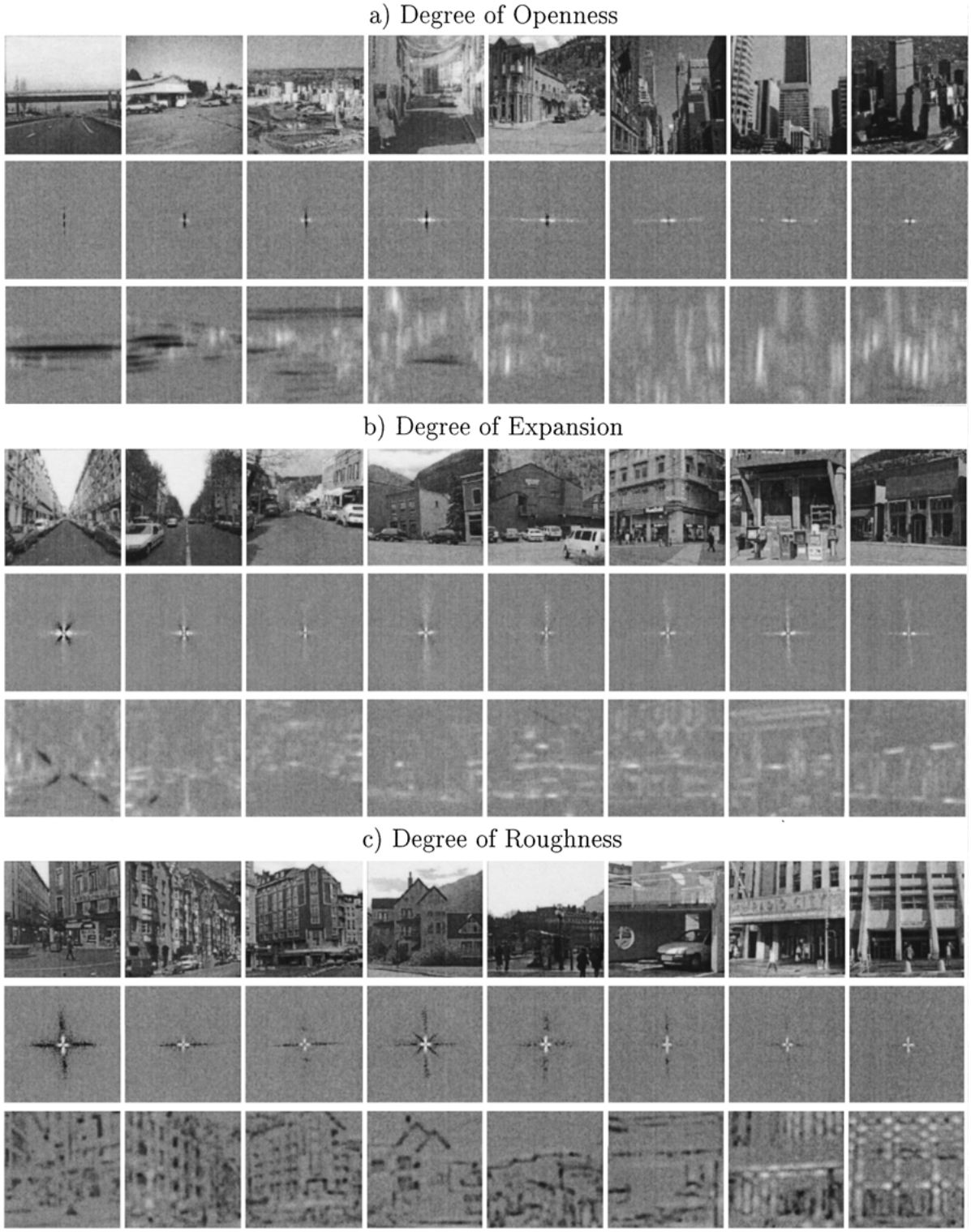

Figure 13. Samples of man-made scene pictures selected at random and ordered with respect to their the degree of openness, degree of expansion, and degree of roughness.

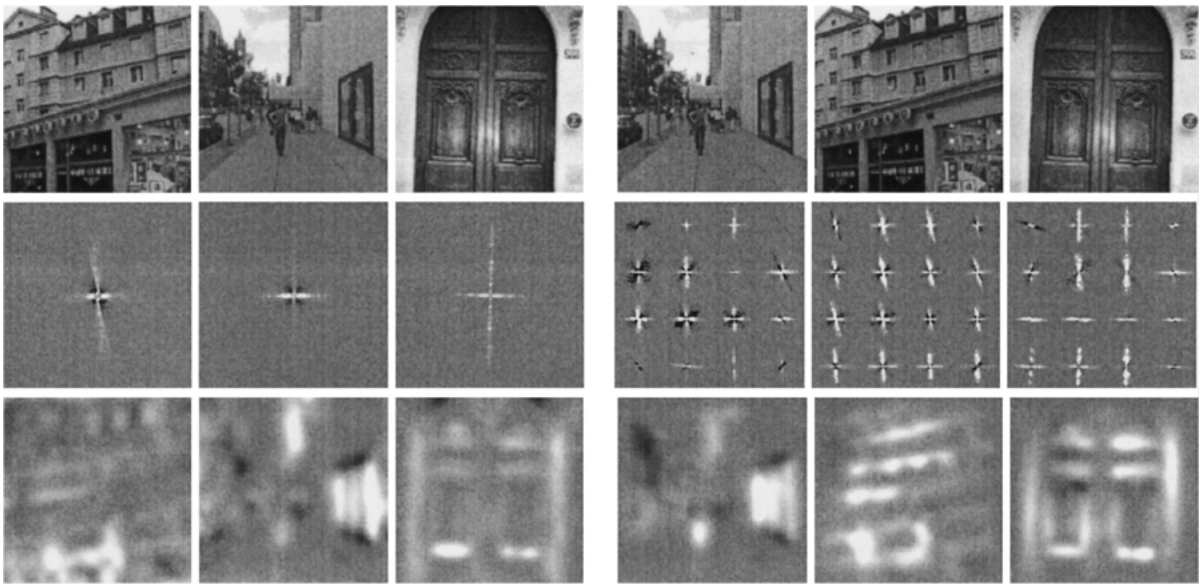

Figure 14. Pictures organized according to the expansion axis. The expansion attribute is estimated by means of the energy spectrum with the DST (left), or by the spectrogram with the WDST (right).

spatial distribution of horizontal and vertical structures in open (Fig. 6(a)) and closed (Fig. 6(b)) man-made environments.

The DST corresponding to the degree of expansion differentiates between diagonal spectral components (in black), corresponding to vanishing lines in the scene, and vertical spectral components and low frequency horizontal components (see Fig. 10(d)). Classification rate of perspective views vs. flat scenes was 90% with the DST (Fig. 10(d)) and 94% with the WDST (Fig. 11(e)). To illustrate the spectral evolution along this axis, we projected a random set of city center scenes (Fig. 13(b)). They were ordered by the DST from scenes with a long perspective (street) to flat views over buildings or houses. Although the local shape of the WDST corresponds roughly to the global DST, diagonal spectral components correlated with a perspective view are not locally symmetrical. Figure 14 illustrates the importance of encoding the spatial configuration of structures. The figure shows three urban scenes organized according to the expansion attribute estimated by both the DST (Fig. 14.left) and the WDST (Fig. 14.right). The DST procedure, which only considers the dominant orientations regardless of location, produced one permutation in the organization due to the abundance of oblique orientations in the left-hand image (the front view of a building). It assigned to that image a high degree of expansion whereas the global pattern of orientations is not in agreement with a strong perspective view. By contrast, the WDST captured the spatial information correlated with the expansion property and considers the building view as less expanded than the sidewalk, but more than the door.

Finally, the roughness DST (see Fig. 10(f)) shows that the two groups differ mainly in their spectral slope along the cardinal orientations. Interestingly, the WDST (Fig. 11(f)) looks similar at each location in the image, independently of the height in the scene (such as the roughness WDST computed for natural scenes). Figure 13(c) illustrates an ordering performed by the roughness DST. Classification rate of new scenes was 91% with the DST and 92% with the WDST. Finally, the correlations between the different templates and human observers are shown in Table 3. The procedure was identical to the natural scenes group procedure. The mean correlation between observers and the DST

Table 3. Correlation between orderings of urban scenes made by observers and the two templates for each spatial envelope property.

	Openness	Expansion	Roughness
Energy spectrum	$m = 0.87$	0.77	0.83
Spectrogram	$m = 0.90$	0.88	0.85
Agreement	0.92	0.91	0.88

Agreement measures the concordance between subjects.

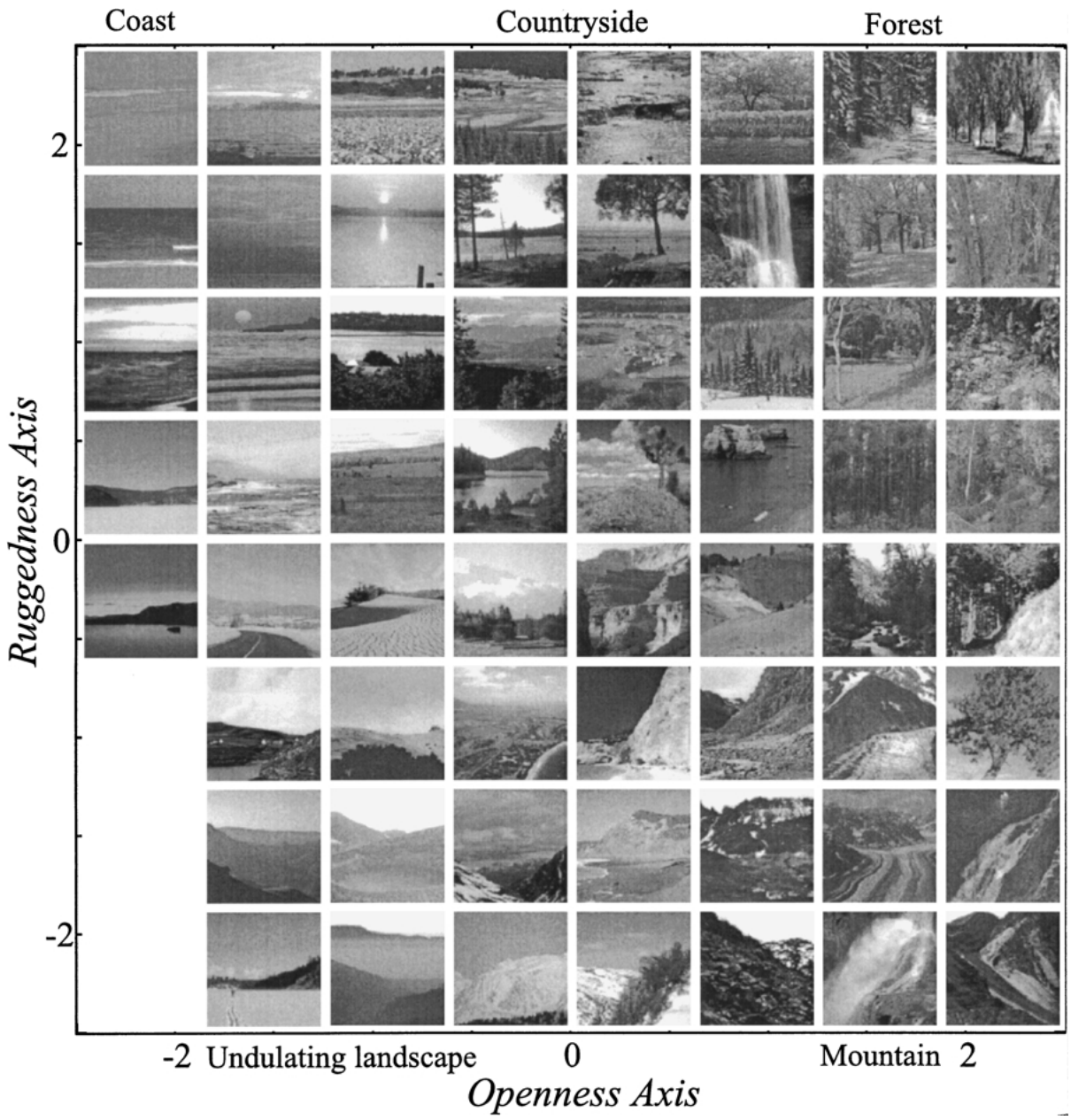

Figure 15. Organization of natural scenes according to the openness and ruggedness properties estimated by the WDSTs.

ranking was 0.82. When using the WDST the rank correlation was 0.87 that was close to the agreement among observers (0.90).

VI. Experimental Results

Each spatial envelope property corresponds to the axes of a multidimensional space into which scenes with similar spatial envelopes are projected closed together. Figures 15 and 16 show a random set of pictures of natural and man-made environments respectively projected in a two dimensional space corresponding to the openness and ruggedness (or expansion for man-made environments) dimensions. Therefore, scenes closed in the space should have the same (or very similar) membership category, whether the spatial envelope

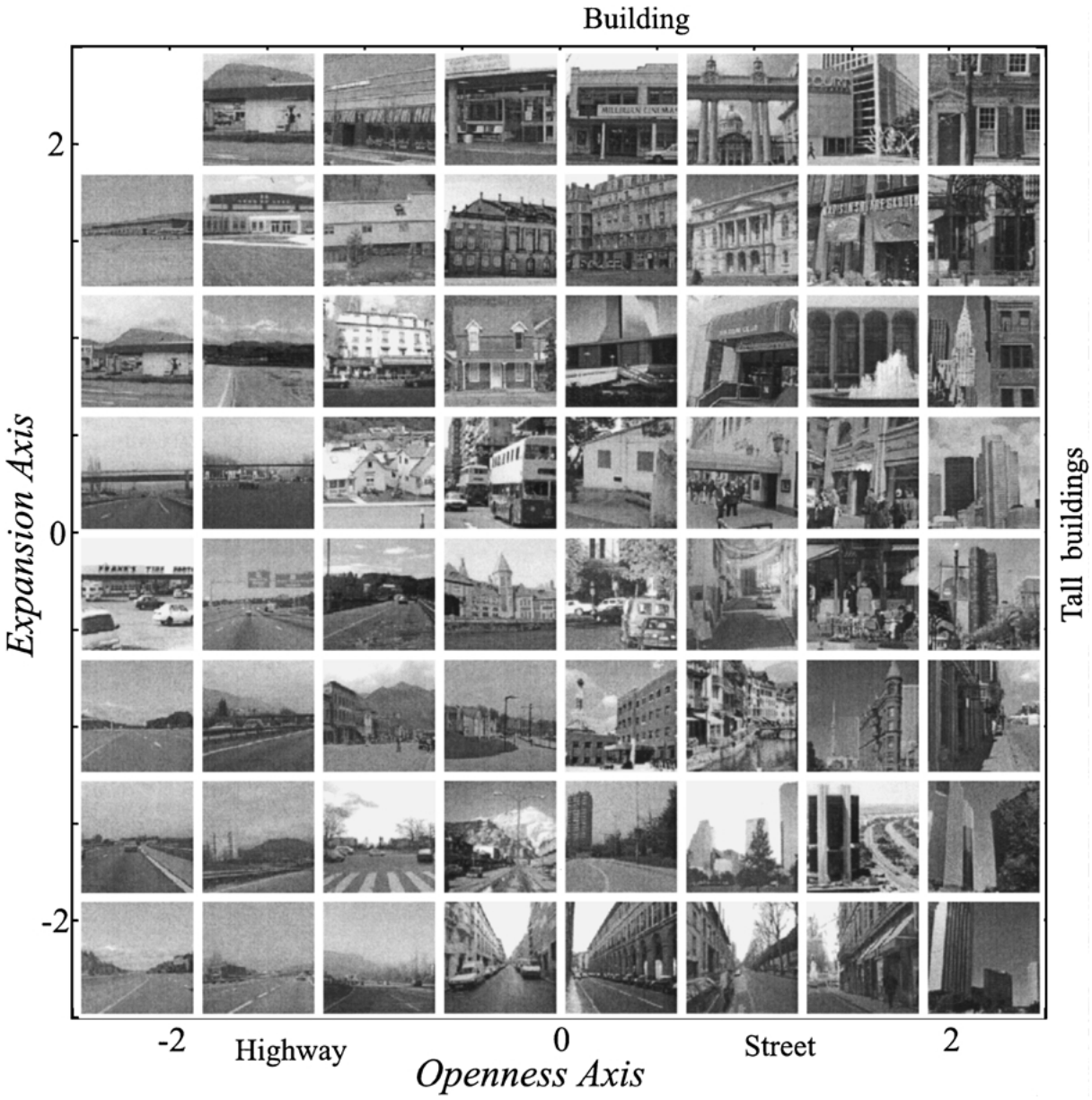

Figure 16. Organization of man-made environments according to the degrees of openness and expansion (WDST).

representation is meaningful enough to provide the semantic category of the scene. We tested this hypothesis as follows: 400 target images out of the image database were chosen at random with their first 7 neighbors. Four subjects were asked to select among the seven neighbors, the scenes that belonged to the same semantic category. For each spatial envelope attribute, similarity between the target scene and each neighbor was approximated by the Euclidean distance between the three attributes (openness, ruggedness/expansion and roughness) of the two images i and j:

$$D^2(i,j) = \left(s^i_{open} - s^j_{open}\right)^2 + \left(s^i_{rugg/exp} - s^j_{rugg/exp}\right)^2 + \left(s^i_{rough} - s^j_{rough}\right)^2 \quad (18)$$

The attributes were normalized in order to have null mean and a standard deviation of 1.

Figure 17. Examples of man-made scenes (target) with four neighbors sharing similar spatial envelope, estimated with the DST and the WDST procedures. The bottom example is an error.

Scenes were considered as correctly recognized when subjects selected at least 4 neighbors as having the same category membership. Figures 17 and 18 show examples of several target scenes with their nearest neighbors (examples of errors are shown at the bottom of both figures). Results of average correct categorization, summarized in Fig. 19, show that, on average among natural and urban landscapes, 92% of the scenes were accurately classified with the WDST and 86% when using the DST. These results highlight

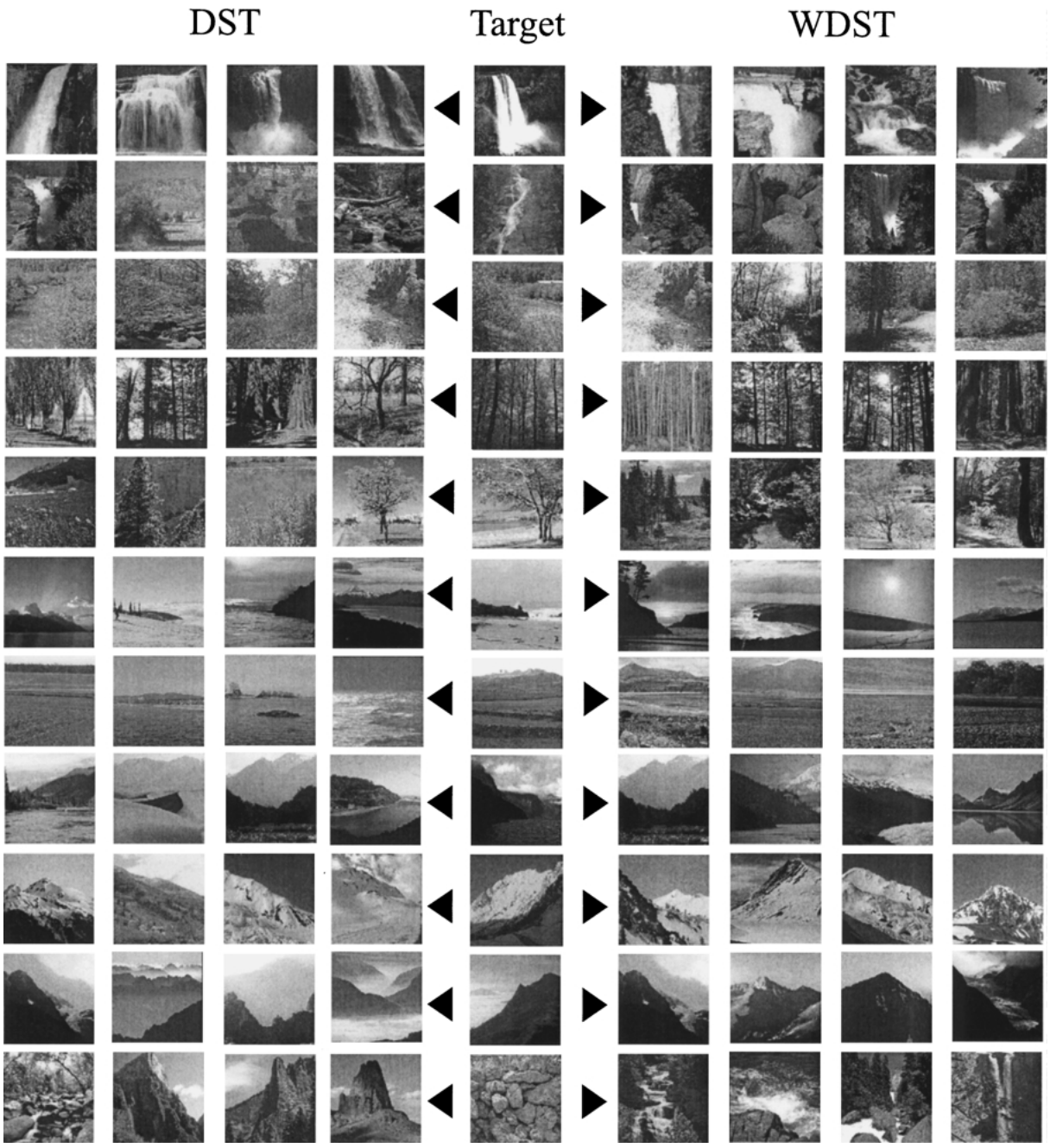

Figure 18. Examples of natural scenes with four neighbors sharing similar spatial envelope, estimated with the DST and the WDST procedures.

the important role played by the unlocalized spectral components (DST) for representing the spatial envelope properties. The addition of spatial layout information clearly increases performance, but most of this performance level may be attributable to the global distribution of the relevant spectral features. In fact, performance differences between the DST and the WDST procedure mostly lie in the *quality* of the spatial envelope neighborhood. In Figs. 17 and 18, we can see that while both DST and WDST models organize together scenes belonging to the same category, the WDST model provides neighbor

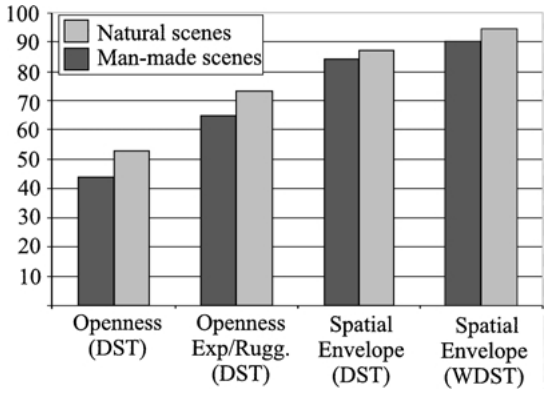

Figure 19. Performances of correct categorization, averaged among subjects, using 200 scenes per group. For a purpose of comparison, it is also shown results of categorization on the same pictures when using the openness DST alone, and the openness with the expanded (or ruggedness) DSTs.

images that are more visually similar to the target scene.

The local similarities between neighbor scenes should be reflected at a global level in the organization of scenes into semantic zones. In Figs. 15 and 16, we can identify a few semantic zones, albeit the organization of pictures is clearly continuous from one border of the space to another. In order to verify the emergence of semantic groups, we projected typical exemplars of coasts, countryside scenes (as fields, open landscapes), forests, mountains, for the natural space, and highways, tall buildings/skyscrapers, streets and close-up city center views for the urban space. Specifically, four observers were asked to choose among the database examples corresponding to these eight categories with the constraint that all the exemplars from a category should also represent visually similar environments. We only kept for the test, the scenes for which three observers agree.

Classification was performed by a K nearest neighbors classifier (K-NN). The K-NN uses a database of labeled scenes with the corresponding category

Table 4. Confusion matrix (in percent) between typical scenes of coasts, countryside (fields, valleys, hills, rolling countryside), enclosed forests and mountains ($N = 1500$).

	Coast	Country	Forest	Mountain
Coast	88.6	8.9	1.2	1.3
Country	9.8	85.2	3.7	1.3
Forest	0.4	3.6	91.5	4.5
Mountain	0.4	4.6	3.8	91.2

Table 5. Confusion matrix (in percent) for the classification between highways, city center streets, city center close views, and tall buildings/skyscrapers ($N = 1400$ images).

	Highway	Street	Close-up	Tall building
Highway	91.6	4.8	2.7	0.9
Street	4.7	89.6	1.8	3.9
Close-up	2.5	2.3	87.8	7.4
Tall building	0.1	3.4	8.5	88

(training set) in order to classify new unlabeled pictures: given a new scene picture, the K-NN first looks for the K nearest neighbors of the new image within the training database. The K neighbors correspond to the K scene pictures from the training database with the smallest distance to the unlabeled picture using the spatial envelope properties (Eq. (18)). Then, it assigns to the new picture the label of the category the most represented within the K nearest neighbors. Performances of classification using the K-NN are presented in Tables 4 and 5 (see Fig. 20 for a sample of misclassified scenes). Classification accuracy was on average 89% with the WDST (and 86% when using the DST procedure). Figure 21 illustrates how the attributes of the spatial envelope differentiate the eight scene categories.

The advantage of representing a scene picture with the spatial envelope attributes is the possibility to

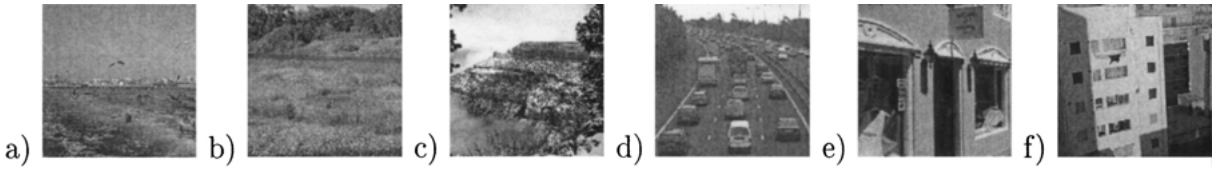

Figure 20. Examples of misclassify scenes using the WDST representation: a) a coast classified as a countryside, b) a field and c) a mountain classified as forests, d) an highway matched with streets, e) a close view classified with tall buildings and f) a tall building classified with close-up views.

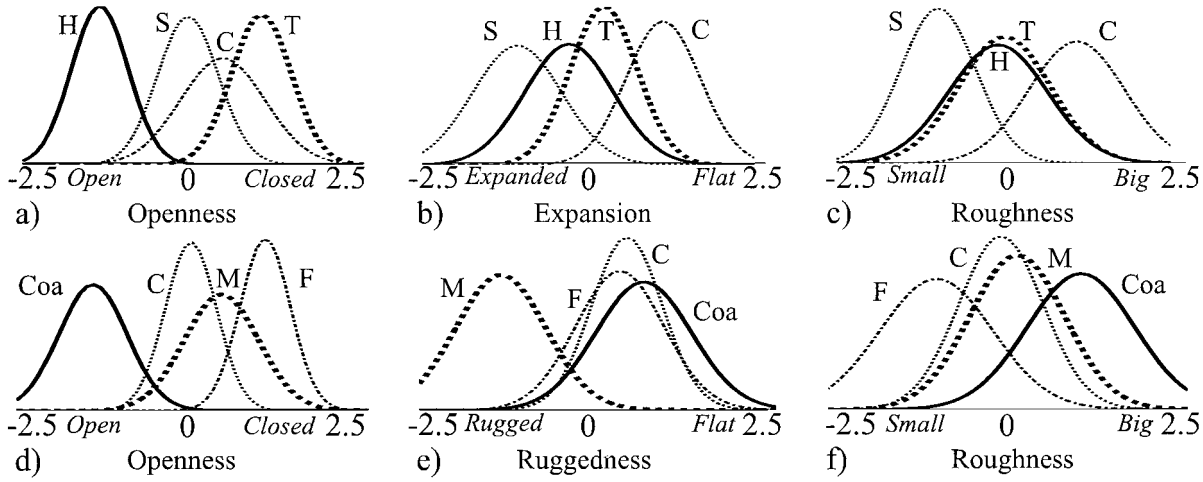

Figure 21. Distribution of man-made (a, b and c) and natural (d, e and f) scene categories along the three attributes of the spatial envelope. Man-made categories: H = highways, S = streets, C = close-up views and T = tall buildings. For natural categories: Coa = Coasts, C = countryside, F = forests and M = mountainous landscapes.

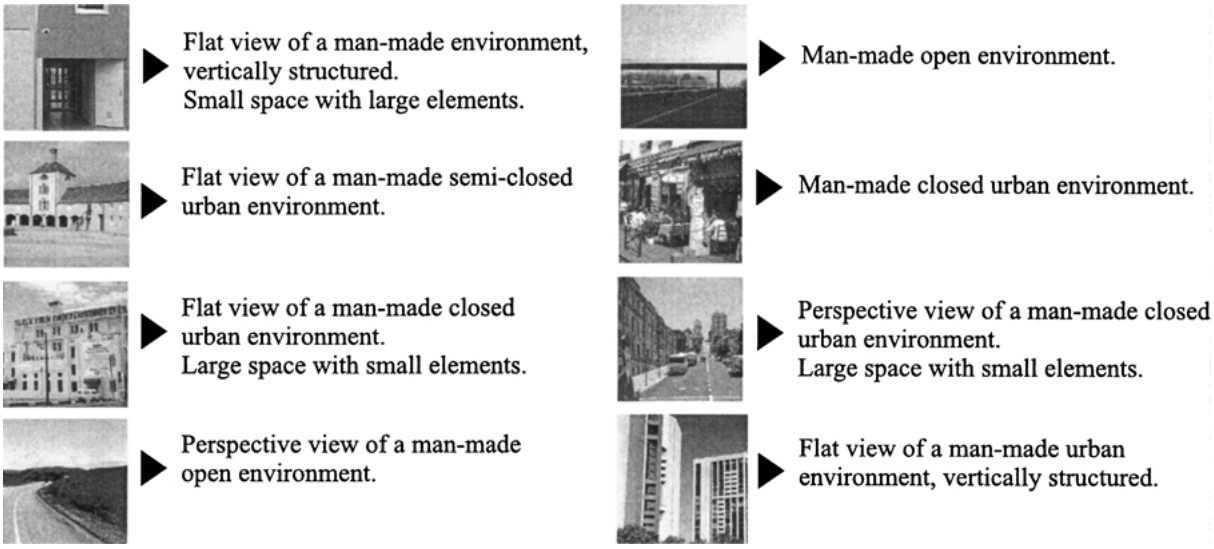

Figure 22. Examples of description of the spatial envelope of urban scenes.

generate a meaningful description of the space that the scene represents. Figure 22 illustrates several examples of descriptions of urban environments, automatically generated. Along the openness dimension, a man-made urban space may be labeled as open, semi-open, closed or vertically structured. Along the two other dimensions, an urban scene can represent a flat vs. expanded view, and a small vs. large space. The centers of the expanded and roughness axes correspond to ambiguous regions, where the system may decide to not attribute a description (Fig. 21).

To summarize, the performances of the two spatial envelope models (based on the DST or the WDST) corroborate the hypothesis that modeling a few structural and spatial properties of natural and urban environmental scenes provide enough information to represent their probable semantic category. The use of holistic scene representations (based on the energy spectrum or the spectrogram) provides a simple procedure for modeling the scene structure that is correlated with semantic aspects. The whole procedure can be summarized in four steps: A) prefiltering of the scene picture, B)

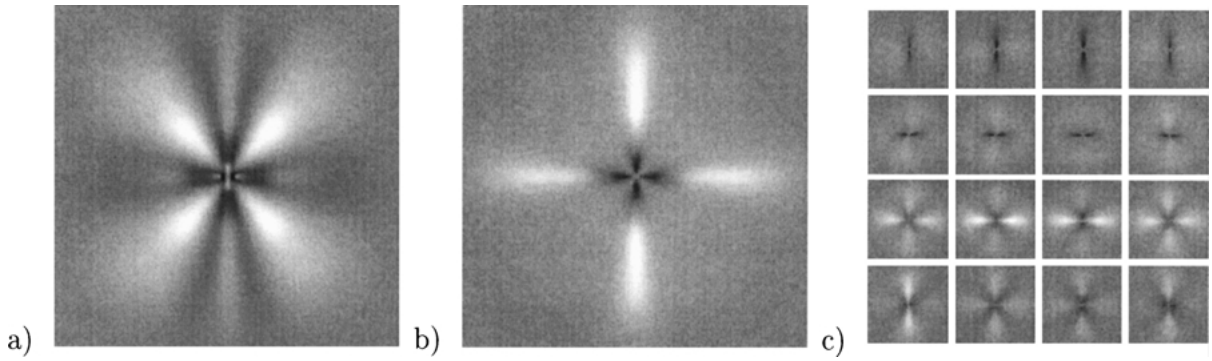

Figure 23. Indoor vs. outdoor DST a) computed on the entire database, and b) computed only with man-made scenes. Dark components correspond to indoor components. C) WDST considering only man-made scenes.

computation of the spectrogram of the filtered image (Eq. (2)), C) classification of the environment into man-made or natural and D) estimation of the three other corresponding spatial envelope attributes with the WDSTs (Eq. (10)). Therefore, these attributes can be used for computing similarities between environmental scenes (Eq. (18)), for generating a meaningful description of the space of the scene or for inferring the probable scene category.

VII. Related Developments

A. *Indoor vs. Outdoor Classification*

Another important issue in scene semantic classification is the distinction between outdoor and indoor images. Separation between these two classes using a global DST (see Fig. 23(a), when considering about 1300 images per group) yields a classification rate of 82%. In fact, the DST is doing a separation similar to the natural vs. man-made scenes separation, as all natural scenes are outdoor images. When dealing with only man-made environments, classification accuracy decreases to 75%. Figure 23(b) shows that the resulting DST looks very similar to the roughness DST. The DST differentiates between indoor scenes that usually contain large objects with flat surfaces and outdoor scenes usually more textured and made of small elements. Using more complex classifiers, as K-NN or Bayesian classifiers (mixture of gaussians), does not provide better results than those obtained by the linear discriminant analysis (DST). These poor results show that global structural features (energy spectrum) are not sufficient to provide reliable performances. In fact, indoor and outdoor environments share similar spectral signatures: they are both made with square-like building blocks based on variations along vertical and horizontal dimensions. Indoor scenes usually have a lower degree of expansion and roughness (the mean slope is about $\alpha \sim 2.5$ for indoors and $\alpha \sim 2.2$ for man-made outdoors), but these characteristics may also be found in some outdoor environments (e.g. front views on a building and outdoors close-up views).

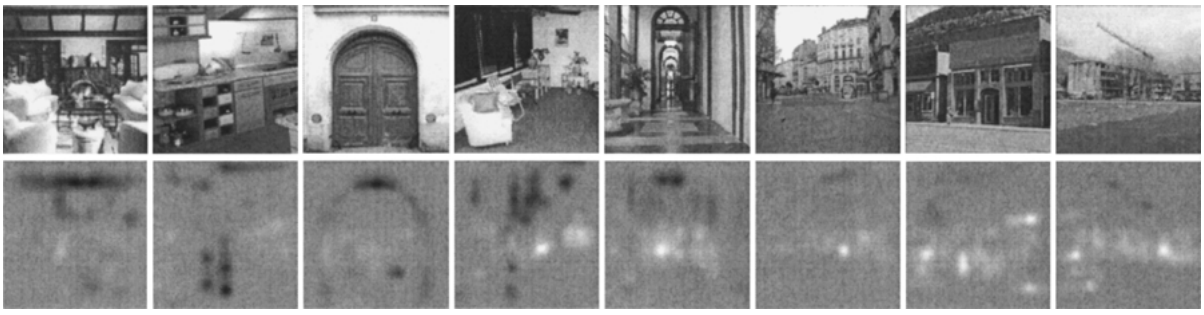

Figure 24. Man-made scenes sorted according to the matching with the WDST for discriminating indoor and outdoor scenes. The decision threshold corresponds to the center. In this example there are two misclassified images.

The spatial information introduced by the spectrogram improves the discrimination. Figure 24 shows a selection of man-made scenes sorted according to the indoor-outdoor WDST (see Fig. 23(c), the WDST was computing with the linear discriminant analysis). The template indicates that indoor scenes are characterized by low spatial frequencies at the bottom of the scene (mainly due to large surfaces), by vertical structures (wall edges, tall furniture, etc.) in the central-top part, and by horizontal structures at the top (e.g. ceiling). Outdoor scenes are mainly characterized by high spatial frequencies everywhere. The WDST procedure gives a 79% of correct classification when considering only man-made structures. When using the decomposition coefficients of the spectrogram, a K-NN classifier ($K = 19$ and $N_L = 40$) provides better results than the WDST procedure (classification accuracy of about 82% when using only man-made scenes, and 85% when considering the whole database).

In any case, an accuracy below 90% is not considered as satisfactory as the indoor vs. outdoor separation is an exclusive choice that does not have any ambiguity for human observers. Current studies devoted to the indoor vs. outdoor classification issue (Szummer and Picard, 1998; Vailaya et al., 1999) have obtained similar results when computing global and local structural attributes (between 80% and 85%). They have already shown that the addition of color cues greatly improve performances over 90%.

B. Depth Perception

Representation of depth in a real world environment is an essential attribute of its semantic representation. When the only information available is the 2D image, the 3D structure of the scene can be partially recovered with algorithms such as shape from shading, from textures, from edges, etc. The spatial envelope may be considered as a simple indicator for the global depth of an environmental scene and may be used as a primary stage for object detection algorithms (Torralba Sinha, 2001). Global depth refers to a global measurement of the mean distance between the observer and the background and main structures inside the scene. Within one set of pictures sharing the same spatial envelope, we observe similar global depths (see Figs. 18 and 17). Scene organizations achieved by the attributes proposed in this paper appear to be correlated with organizations based on global depth (Figs. 12 and 13).

C. Environmental Scenes vs. Object Discrimination

This paper focuses on the definition of meaningful attributes for differentiating among environmental scenes. However, in a real setting, many of the views that a system has to deal with when interacting with its environment will correspond to close-up views of objects and structures (tools, faces, furniture, etc.). Therefore, a remaining question to explore is whether there exist structural and configuration differences between environmental scenes (indoor and outdoor) and object views that would allow their discrimination. Most man-made object views contain sharp edges and flat surfaces, without a statistical bias with respect to dominant orientations due to the variety in object shapes and the variations of point of views. On the contrary, scenes have few sharp edges and flat surfaces with a strong bias towards the horizontal and vertical orientations. They also have more restrictions in their point of view due to the scale of the space larger than human size. However, as these tendencies are not shared by all the members of the group and environmental scene group, the energy spectrum (DST) provides only a 72% of correct classification.

With respect to the spatial configurations, there are relatively few constraints on object shapes and their spatial arrangements. Given an environmental scene, there are many object arrangements that produce a similar configuration of the spectral features in the image. However, most environmental scenes have strong configuration rules that are rarely found in non-accidental object arrangements (for example, the layout of divergent lines of a perspective view of a street). The spectrogram features, using a WDST (Fig. 25), provide a 82% of correct discrimination between man-made objects and environmental scenes (with 622 pictures for each group). The WDST mainly reveals the discriminant layout between objects and scenes: it assigns to the scenes the configuration of spectral components in agreement with an expanded space (concave space with perspective lines going away from the observer). On the contrary, the spectral components assigned to objects (see an illustration in the opponent energy image of Fig. 26) correspond more to convex structures. Perspective lines coming towards the observer, that are common to many object views, are rarely found in typical views of environmental scenes. The discrimination result of 82% shows that there are significant structural and configurational differences between objects and environmental scenes that can be captured by a

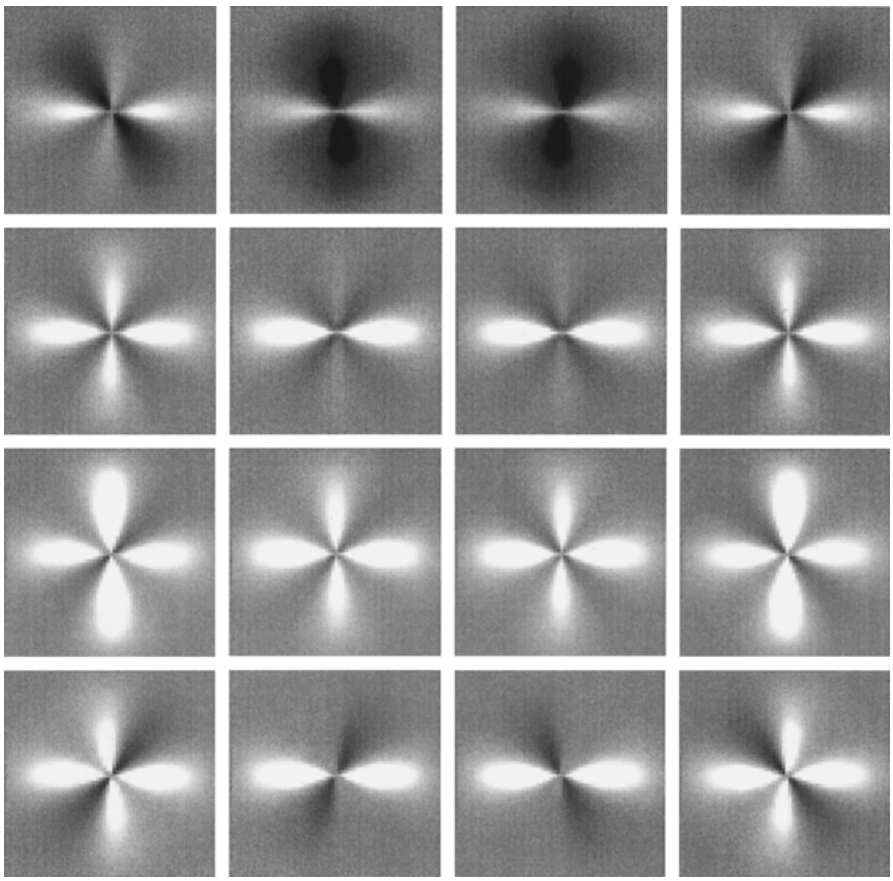

Figure 25. WDST ($N_L = 20$) for classification of man-made environmental scenes and objects. Dark components correspond to close-up views of objects.

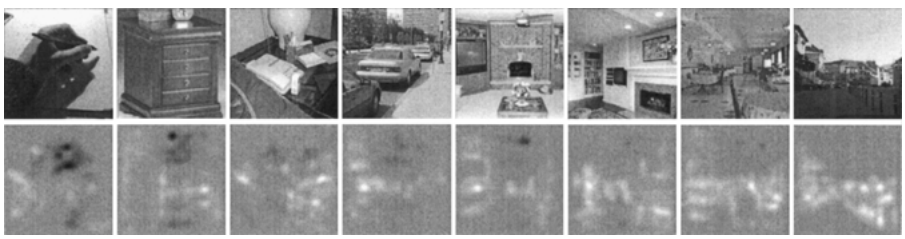

Figure 26. A sample of pictures sorted according to the matching with the object-scene WDST, and their respective opponent energy image. The decision threshold corresponds to the center.

holistic image representation. We should also emphasize that this discrimination task is not as exclusive as the indoor/outdoor discrimination (e.g. human observers usually assign an image representing a house to the "object" group).

VIII. Conclusion

This paper introduces a representation of the structure of real world scenes that we termed *Spatial Envelope*. The spatial envelope of an environmental scene may be described as a set of perceptual properties (naturalness, openness, roughness, ruggedness and expansion). These properties are related to the shape of the space and are meaningful to human observers. The spatial envelope properties provide a holistic description of the scene where local object information is not taken into account.

We show that these spatial properties are strongly correlated with the second-order statistics (DST) and with the spatial arrangement of structures in the scene (WDST). The spatial envelope model organizes scene pictures as human subjects do, and is able to retrieve images that share the same semantic category. The performance of this holistic model corroborates the assumption that object information is not a necessary stage for achieving the scene identity level. Therein, spatial envelope attributes may be used for computing similarities between environmental scenes, for generating a meaningful description of the space that the scene subtends, or for inferring the scene category. It provides a meaningful representation of complex environmental scenes that may sketch a direct interface between visual perception and semantic knowledge. The holistic scene representation provides an efficient way for context modeling that can be used as the first stage of object detection algorithms by priming typical objects, scales and locations (Torralba and Sinha, 2001).

Appendix I: Principal Components of the Energy Spectrum and Spectrogram

For images of size N^2 pixels, the function energy spectrum, $A(f_x, f_y)^2$, obtained from the Discrete Fourier Transform can be sampled giving N^2 values. Similarly, the spectrogram, $A(x, y, f_x, f_y)^2$, computed at 8×8 spatial locations will give $64N^2$ samples. The goal of PCA is to transform the original set of variables in a reduced number of uncorrelated variables that preserve the maximal variation of the original data (e.g. Moghaddam and Pentland, 97). In order to compute the principal components, we rearrange the samples of the energy spectrum (or spectrogram) in a column vector $\mathbf{p}$. The KL basis for both representations are obtained as the eigenvectors of the covariance matrix $E[(\mathbf{p} - \bar{\mathbf{p}})\,(\mathbf{p} - \bar{\mathbf{p}})^T]$, with $\bar{\mathbf{p}} = E[\mathbf{p}]$. The expectations are computed from averaging across a large image database. Finally, the PCA extracts the subspace spanned by a subset of KL functions with the largest eigenvalues. In practice, the number of training images is $N_i < N^2$ and, therefore, the KL basis cannot be estimated reliably. A way to reduce the dimensionality of the vector $\mathbf{p}$ is to reduce the number of samples of the function $A(f_x, f_y)$, or $A(x, y, f_x, f_y)$. We propose to sample the function $A(f_x, f_y)$ as:

$$g_i = \int\int A(f_x, f_y)^2\, G_i(f_x, f_y)\, df_x\, df_y \qquad (19)$$

G_i are a set of Gaussian functions arranged in a log-polar array and calculated by rotating and scaling the function: $G(f_x, f_y) = e^{-f_y^2/\sigma_y^2}(e^{-(f_x - f_o)^2/\sigma_x^2} + e^{(f_x + f_o)^2/\sigma_x^2})$. In this paper we use a set of Gaussians distributed in 5 frequency bands with central frequencies f_o at 0.02, 0.04, 0.08, 0.16 and 0.32 c/p, and 12 orientations at each band (other configurations yield similar results). The function $A(f_x, f_y)^2$ is then represented by a vector $\mathbf{g} = \{g_i\}_{i=1,L}$ with dimensionality $L < N_i$. The same spectral sampling can be applied locally to the spectrogram providing a vector $\mathbf{g}$ with $64L$ samples when considering 8×8 different spatial locations. The log-polar sampling corresponds to a change of coordinates of the spectrum from Cartesian coordinates (f_x, f_y) to polar coordinates (f, θ). This transformation provides the same detail of description at each spatial scale in agreement with the property of scale invariance of the second order statistics of real world images (e.g. Field, 1987). In our experiments, we have noticed that the log-polar transformation reduces the number of principal components needed for the discrimination of scene categories compared to the original Cartesian coordinates. PCA is then applied to the low dimensional vector $\mathbf{g}$. We compose a new vector $\mathbf{v} = \mathbf{C}\mathbf{g}$ with decorrelated components by solving the matrix equation: $E[\mathbf{v}\,\mathbf{v}^T] = \mathbf{C}E[(\mathbf{g} - \bar{\mathbf{g}})\,(\mathbf{g} - \bar{\mathbf{g}})^T]\mathbf{C}^T = \Lambda$, being Λ a diagonal matrix. The matrix $\mathbf{C} = \{c_{i,j}\}$ contains the eigenvectors of the covariance matrix and the diagonal matrix Λ contains the associated eigenvalues sorted in decreasing order. We select only the N_G first components of the vector $\mathbf{v} = \{v_i\}_{i=1,L}$. When the

vector **g** is obtained from the global energy spectrum, the components v_i are then computed as:

$$v_i = \sum_{j=1}^{L} c_{i,j} g_j = \int\int A(f_x, f_y)^2 \, \psi_i(f_x, f_y) \, df_x \, df_y \tag{20}$$

with:

$$\psi_i(f_x, f_y) = \sum_{j=1}^{L} c_{i,j} G_j(f_x, f_y) \quad i = 1, \, N_G \tag{21}$$

The features v_i are decorrelated. The functions $\psi_i(f_x, f_y)$ are the approximations of the N_G principal KL functions. A similar derivation applies when the vector **g** is obtained from the spectrogram.

Acknowledgments

The authors would especially like to thank W. Richards and P. Sinha for fruitful discussions, and two anonymous reviewers whose comments greatly helped improve the manuscript. Thanks also to Megan Hyle for useful comments about this work.

Notes

1. The "gist" is an abstract representation of the scene that spontaneously activates memory representations of scene categories (a city, a mountain, etc.) (see Friedman, 1979).
2. Looking for a more precise nomenclature describing the openness of a scene, we found early suggestions given by Gibson (1976), who defines an open environment as the "layout consisting of the surface of the earth alone" (1986, p. 33). Then, surfaces of the ground are usually more or less "wrinkled by convexities and concavities".
3. We apply a pre-filtering to the input image $i(x, y)$ that reduces illumination effects and prevents some local image regions to dominate the energy spectrum. The prefiltering consists in a local normalization of the intensity variance:

   $$i'(x, y) = \frac{i(x, y) * h(x, y)}{\epsilon + \sqrt{[i(x, y) * h(x, y)]^2 * g(x, y)}}$$

   $g(x, y)$ is an isotropic low-pass gaussian spatial filter with a radial cut-off frequency at 0.015 cycles/pixel, and $h(x, y) = 1 - g(x, y)$. The numerator is a high-pass filter that cancels the mean intensity value of the image and whitens the energy spectrum at the very low spatial frequencies. The denominator acts as a local estimator of the variance of the output of the high-pass filter. ϵ is a constant that avoids noise enhancement in constant image regions. We set experimentally $\epsilon = 20$ for input images with intensity values in the range $[0, 255]$. This pre-filtering stage affects only the very low spatial frequencies (below 0.015 c/p) and does not change the mean spectral signature.
4. The naturalness DST coefficients do not vary, neither performances of categorization, when more images are used for the learning stage.
5. Note that in this representation of the two filters, high spatial frequencies hide the contribution of low spatial frequencies, which is not the case when looking at the DST in the spectral domain.
6. The same templates may be computed using the linear discriminant analysis (see Torralba and Oliva, 1999) with two scene groups that are exclusive along each property. By doing so, we have obtained templates similar to those shows in Figs. 10 and 11, and similar performances in classification and ordering.

References

Amadasun, M. 1989. Textural features corresponding to textural properties. *IEEE Trans. Sys., Man and Cybernetics*, 19:1264–1274.

Atick, J. and Redlich, A. 1992. What does the retina know about natural scenes? *Neural Computation*, 4:196–210.

Baddeley, R. 1997. The correlational structure of natural images and the calibration of spatial representations. *Cognitive Science*, 21:351–372.

Barrow, H.G. and Tannenbaum, J.M. 1978. Recovering intrinsec scene characteristics from images. In *Computer Vision Systems*, A. Hanson and E. Riseman (Eds.), Academic Press: New York, pp. 3–26.

Biederman, I. 1987. Recognition-by-components: A theory of human image interpretation. *Psychological Review*, 94:115–148.

Biederman, I. 1988. Aspects and extension of a theory of human image understanding. In *Computational Processes in Human Vision: An Interdisciplinary Perspective*, Z. Pylyshyn (Ed.), Ablex Publishing Corporation: Norwood, New Jersey.

Carson, C., Belongie, S., Greenspan, H., and Malik, J. 1997. Region-based image querying. In *Proc. IEEE W. on Content-Based Access of Image and Video Libraries*, pp. 42–49.

Carson, C., Thomas, M., Belongie, S., Hellerstein, J.M., and Malik, J. 1999. Blobworld: A system for region-based image indexing and retrieval. In *Third Int. Conf. on Visual Information Systems*, June 1999, Springer-Verlag.

De Bonet, J.S. and Viola, P. 1997. Structure driven image database retrieval. *Advances in Neural Information Processing*, 10:866–872.

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.

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

Field, D.J. 1994. What is the goal of sensory coding? *Neural Computation*, 6:559–601.

Friedman, A. 1979. Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, 108:316–355.

Guerin-Dugue, A. and Oliva, A. 2000. Classification of scene photographs from local orientations features. *Pattern Recognition Letters*, 21:1135–1140.

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

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

Heaps, C. and Handel, S. 1999. Similarity and features of natural textures. *Journal of Experimental Psychology: Human Perception and Performance*, 25:299–320.

Henderson, J.M. and Hollingworth, A. 1999. High level scene perception. *Annual Review of Psychology*, 50:243–271.

Hochberg, J.E. 1968. In the mind's eye. In *Contemporary Theory and Research in Visual Perception*, R.N. Haber (Ed.), Holt, Rinehart, and Winston: New York, pp. 309–331.

Lipson, P., Grimson, E., and Sinha, P. 1997. Configuration based scene classification and image indexing. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Puerto Rico, pp. 1007–1013.

Marr, D. 1982. *Vision*. WH Freeman: San Francisco, CA.

Moghaddam, B. and Pentland, A. 1997. Probabilistic Visual Learning for Object Representation. *IEEE Trans. Pattern Analysis and Machine Vision*, 19(7):696–710.

Morgan, M.J., Ross, J., and Hayes, A. 1991. The relative importance of local phase and local amplitude in patchwise image reconstruction. *Biological Cybernetics*, 65:113–119.

Oliva, A. and Schyns, P.G. 1997. Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, 34:72–107.

Oliva, A. and Schyns, P.G. 2000. Diagnostic color blobs mediate scene recognition. *Cognitive Psychology*, 41:176–210.

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

O'Regan, J.K., Rensink, R.A., and Clark, J.J. 1999. Change-blindness as a result of 'mudsplashes'. *Nature*, 398:34.

Piotrowski, L.N. and Campbell, F.W. 1982. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11:337–346.

Pentland, A.P. 1984. Fractal-based description of natural scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:661–674.

Potter, M.C. 1975. Meaning in visual search. *Science*, 187:965–966.

Rao, A.R. and Lohse, G.L. 1993. Identifying high level features of texture perception. *Graphical Models and Image Processing*, 55:218–233.

Rensink, R.A. 2000. The dynamic representation of scenes. *Visual Cognition*, 7:17–42.

Rensink, R.A., O'Regan, J.K., and Clark, J.J. 1997. To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science*, 8:368–373.

Ripley, B.D. 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK.

Rosch, E. and Mervis, C.B. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

Sanocki, T. and Epstein, W. 1997. Priming spatial layout of scenes. *Psychological Science*, 8:374–378.

Sanocki, T. and Reynolds, S. 2000. Does figural goodness influence the processing and representation of spatial layout. *Investigative Ophthalmology and Visual Science*, 41:723.

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.

Simons, D.J. and Levin, D.T. 1997. Change blindness. *Trends in Cognitive Sciences*, 1:261–267.

Sirovich, L. and Kirby, M. 1987. Low-dimensional procedure for the characterization of human faces. *Journal of Optical Society of America*, 4:519–524.

Swets, D.L. and Weng, J.J. 1996. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 18:831–836.

Switkes, E., Mayer, M.J., and Sloan, J.A. 1978. Spatial frequency analysis of the visual environment: anisotropy and the carpentered environment hypothesis. *Vision Research*, 18:1393–1399.

Szummer, M. and Picard, R.W. 1998. Indoor-outdoor image classification. In *IEEE intl. Workshop on Content-Based Access of Image and Video Databases*.

Tamura, H., Mori, S., and Yamawaki, T. 1978. Textural features corresponding to visual perception. *IEEE Trans. Sys. Man and Cybernetics*, 8:460–473.

Torralba, A. and Oliva, A. 1999. Scene organization using discriminant structural templates. In *IEEE Proc. Of Int. Conf in Comp. Vision*, pp. 1253–1258.

Torralba, A. and Oliva, A. 2001. Depth perception from familiar structure. submitted.

Torralba, A. and Sinha, P. 2001. Statistical context priming for object detection. In *IEEE. Proc of Int. Conf. in Computer Vision*.

Tversky, B. and Hemenway, K. 1983. Categories of environmental scenes. *Cognitive Psychology*, 15:121–149.

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

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