Semantic Organization of Scenes using Discriminant Structural Templates

Antonio Torralba and Aude Oliva
LIS-INPG, Grenoble, France.
torralba@ai.mit.edu, oliva@search.bwh.harvard.edu

*Proceedings of International Conference on Computer Vision, ICCV99,
Korfu, Grece, September 1999, pages 1253-1258.*

Abstract

In this paper, we present a procedure for organizing real world scenes along semantic axes. The approach is based on the output energies of linear discriminant filters that take into account, or not, spatial information.

We introduce three semantic axes along which pictures are ordered. The main semantic axis computes the degree of naturalness of a scene. Then, urban pictures are evaluated according to their degree of verticalness and natural scenes, according to their degree of openness. We observe the emergence of typical scene categories such as beach, mountain, skyscrapers, city center, etc., along the axes.

1 Introduction

Human observers recognize complex visual scenes in a single glance, in spite of the numerous of objects they contain, with different colors, shadows, textures, etc. To resolve it, the visual system automatically extracts a global information about the main structure of the scene, ignoring most of details and objects information [4-6].

In this paper, we introduce a computational procedure that extracts a global structural information from complex scenes. Common to recent studies about scene recognition [1-4,10-11] is the classification into exclusive classes. However, when dealing with a very large database, exclusive classification may increase irrelevant classification rate as most of scenes are ambiguous in terms of category. Our approach proposes several structural attributes that allow to organize continuously scenes along semantic axes [7]. The structural attributes are computed from the output energies of linear filters. By computing a global structural attribute for each scene (e.g. a city skyline is vertically structured, a coast is horizontally structured), we observe that scenes belonging to the same category (e.g. city center, skyscraper, forest, mountain, etc.) are grouped together whereas ambiguous scenes in terms of category (tall buildings in a center area; rocky valley with trees) are located between semantic zones.

To explore computation of the ”main structure” of a scene, the next sections details two kinds of optimal filters: a global filter computed over the whole image and a spatial variant filter.

2 Semantic Axes

This paper details two levels of semantic axes that represents attributes of the main structure of a scene:

1) The first semantic axis represents the *degree of naturalness* of a scene. This axis goes from man-made environments to natural landscapes. Ambiguous pictures in terms of ”artificiality” (as a farm in a field) are likely to be projected around the center of the axis.

2) The second semantic axis depends on the organization provided by the first axis. Natural scenes are represented according to their *degree of openness*, from panoramic scenes (e.g. coast, beach) to closed environments (e.g. forest, mountain). Degree of openness of artificial urban scenes is estimated according to their quantity of horizontal and vertical lines. This axis goes from vertical to horizontally dominant scenes (from highways to tall buildings).

3 Computation of structural attributes

Main orientations and spatial frequency distributions of the main structure of a scene are encoded in its power spectrum. We show that the power spectrum contains relevant information for the evaluation of the semantic axes.

3.1 Image Power Spectrum

The power spectrum of an image is computed by taking the squared magnitude of its Fourier Transform:

$$\Gamma(f_x, f_y) = |FT\{i(x,y)\}|^2 \qquad (1)$$

where $i(x, y)$ is the intensity distribution of the image along the spatial variables x and y. FT is the Fourier Transform, f_x and f_y are the spatial frequencies. Power spectrum, $\Gamma(f_x, f_y)$, encodes the energy density for each spatial frequency and orientations over the whole image.

3.2 Discriminant Spectral Templates

A Discriminant Spectral Template (DST) is represented by a set of low-level features (here orientation and spatial frequency distributions) encoding the structure which is discriminant between two scene categories. For example, panoramic scenes versus textured scenes (forests) are discriminated by opposing vertical spectral components (axe f_y) versus other orientations, see Fig. 1.b.

A structural discriminant feature, u, is computed per picture by using a DST as follows:

$$u = \iint \Gamma(f_x, f_y)\, DST(f_x, f_y) df_x\, df_y \qquad (2)$$

u is a weighted integral of the power spectrum of the image. $DST(f_x, f_y)$ is the weighted function that describes how each spectral component contributes to the structural attribute.

Fig. 1 shows several examples of $DSTs$ that extract structural attributes. The dark and white pixels correspond respectively to the negative and positive values. These graphical representations allow an easy understanding of the way that scene organization is performed: as an illustration, look at the Naturalness DST (Fig 1.a). Artificial components are represented by the dark zones describing a cross form, but only at medium and high spatial frequencies. This means that a scene picture will be labeled as an artificial environment whether its power spectrum mostly matches the dark zones of the DST. The white part corresponds to a natural structure: it is composed with more oblique elements from low to high spatial frequencies and a vertical spectral component only at low spatial scale (the ”horizon” scenes). Performances of the DSTs are presented in the section 4.

3.3 Learning

A supervised learning stage is used to determine the DST associated to each semantic axis.

In order to represent DSTs in a low dimensional space, we decompose it into a set of functions $G_n(f_x, f_y)$ that do not need to be orthogonal. In this paper, we use gaussian envelopes that correspond to Gabor filters:

$$DST(f_x, f_y) = \sum_{n=1}^{N} d_n\, G_n(f_x, f_y)^2 \qquad (3)$$

The coefficients d_n show how weighting each Gabor filter in order to build $DSTs$. The coefficients d_n will be determined by the learning stage. By replacing eq. (3) into eq. (2), we obtain the next equation:

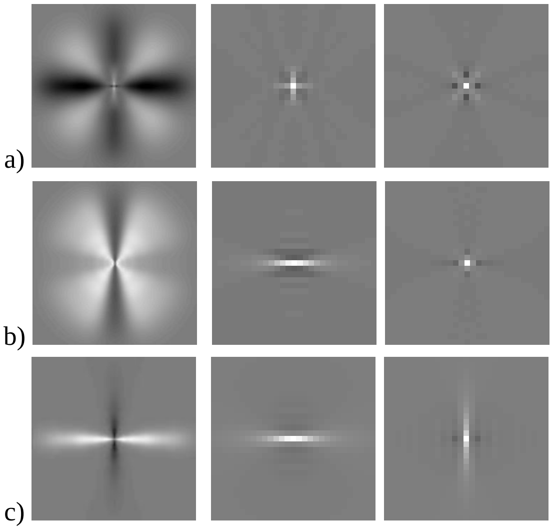

Figure 1. The left-hand column shows the 3 DSTs. The middle and right-hand columns show h_- and h_+. a) artificial vs. natural. b) open vs. closed natural scenes. c) horizontal vs. vertical artificial scenes.

$$g_n = \iint \Gamma(f_x, f_y)\, G_n(f_x, f_y)^2\, df_x\, df_y \qquad (4)$$

where g_n are the output energies for the N Gabor filters used as basis for the DST. We can compute u for each image from these energies as:

$$u = \sum_{n=1}^{N} d_n\, g_n \qquad (5)$$

Here, we sampled the power spectrum with $N = 70$ Gabor filters from high spatial frequencies (1/3 cycles/image) to low spatial frequencies (1/72 cycles/image). But no differences are when using different reasonable values of N.

In the learning step, each image is represented by a vector of features $\mathbf{x} = \{g_n\}$, g_n being the output energies of a set of Gabor filters.

Several methods can be used to determine the coefficients d_n. The method presented here consists in finding two different sets of images that can be described with unambiguous semantic attributes. As an illustration, consider the artificiality of a scene. The first group will be composed of images containing only man-made structures and the second one will contain only natural landscape images. The DST must at best separate these two groups. The parameters of the DST, d_n, can be learnt by applying Linear Discriminant Analysis [8, 9] which looks for the parameters giving the best classification rate. Two matrices are fundamental when applying the discriminant analysis. The covariance matrix: $\mathbf{T} = E[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T]$. $\mathbf{m} = E[\mathbf{x}]$ is the mean vector of features. The between-class scatter matrix is defined as: $\mathbf{T}_b = (\mathbf{m}_1 - \mathbf{m})(\mathbf{m}_1 - \mathbf{m})^T + (\mathbf{m}_2 - \mathbf{m})(\mathbf{m}_2 - \mathbf{m})^T$. $\mathbf{m}_1$ and $\mathbf{m}_2$ are the mean vectors of the feature vectors of the two classes. The between-class scatter matrix measures the distance between the centers of the two classes. The discriminant analysis determines the discriminant projection vector that maximizes the distance between the two classes after projection [1]. The discriminant projection vector corresponds to the vector $\mathbf{d} = \{d_n\}$ as defined in the previous section. The discriminant vector is the eigenvector of the matrix $\mathbf{T}^{-1}\,\mathbf{T}_b$ with the largest eigenvalue. As only two groups per semantic axis are defined, only one eigenvalue is different from zero. Therefore, only one discriminant projection vector can be defined. The discriminant projection vector is given by: $\mathbf{d} = \mathbf{T}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$. The inversion of $\mathbf{T}$ may be ill-conditioned when the number of examples is not larger enough. In that case we use classic regularization techniques (principal components or adding a perturbation to the matrix $\mathbf{T}$. There is no difference in the resulting organizations).

Once the learning is done, we compose DSTs using the equation (3). After that, projection of one image into the semantic axis does not require the computation of any Gabor filters. The computational steps for obtaining the structural feature (and, therefore, the position of the image along the semantic axis) are: 1) *Prefiltering*: We divided the image intensity at each pixel by an estimation of the local variance in order to reduce illuminant variations. 2) *Power spectrum computation*. 3) *Structural feature computation*: using equation (2). All this procedure requires only global and simple computations on the image yielding to a very efficient algorithm that gives structural information about the scene. The left-hand side of Fig. 1 shows three global DSTs. The first one organizes images from artificial to natural scenes. The second one organizes natural landscapes from open to closed scenes and the third one organizes artificial areas from horizontally to vertically structured scenes.

3.4 Scene Discriminant Filters

The output energy of a filter with transfer function $H(f_x, f_y)$ can be computed as:

$$E = \int\!\!\int \Gamma(f_x, f_y)\; |H(f_x, f_y)|^2\, df_x\, df_y \qquad (6)$$

This expression is similar to eq. (2) used to compute the structural feature u. However, as the squared magnitude of the transfer function of a filter cannot have negative values, the DST function cannot be implemented by a unique filter. In fact, it can be implemented by computing the difference between the output energies of two filters [2]. In such a case, we can compute u as the difference between two energies as $u = E_+ - E_-$, where E_+ and E_- are respectively the output energy of two filters with transfer functions H_+ and H_-. In such a case, we obtain:

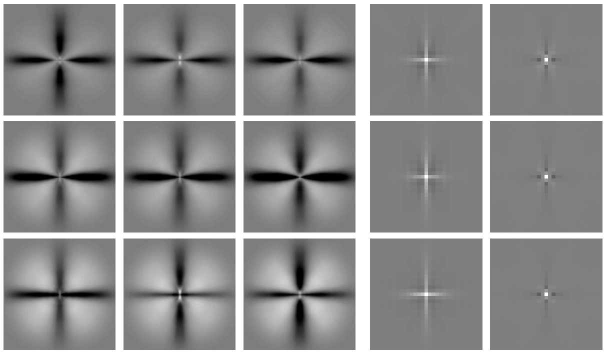

Figure 2. Spatial variant DST for Artificial versus Natural scenes. The 9 templates at the left-hand side are the local DSTs. At the right hand side we show the filters h_- and h_+.

$$u = \int\!\!\int \Gamma(f_x, f_y)\,(|H_+(f_x, f_y)|^2 - |H_-(f_x, f_y)|^2) df_x\, df_y$$

This expression allows us to write the DST as:

$$DST = |H_+|^2 - |H_-|^2 \qquad (7)$$

With this expression, it is possible to obtain positive and negative values for the DST. Several functions H_+ and H_- give the same resulting DST. Here, we use:

$$|H_+(f_x, f_y)|^2 = \sum_{n=0}^{N} p(d_n)\, G_n(f_x, f_y)^2 \qquad (8)$$

and

$$|H_-(f_x, f_y)|^2 = \sum_{n=0}^{N} p(-d_n)\, G_n(f_x, f_y)^2 \qquad (9)$$

where $p(x) = x$ if $x > 0$ and $p(x) = 0$ if $x < 0$. d_n are the components of the discriminant projection vector computed in the learning stage. These equations give the magnitude of the two filters. As the phase can be freely chosen, we chose null phase filters.

[1] Reduction of the data can be performed before applying the discriminant analysis. However such an operation does not improve the results as the standard deviations of the Gabor energy outputs are very similar for the configuration chosen here.

[2] It must be noted that although scene category discrimination is possible using an unique filter, it will yield to poorer results than using two filters. Using more than two filters will not improve results.

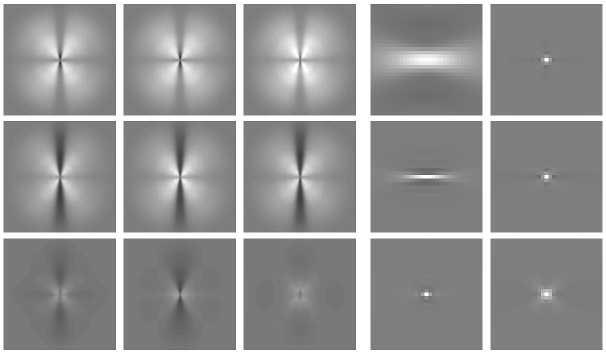

Figure 3. Spatial DST for Open vs. Closed natural scenes.

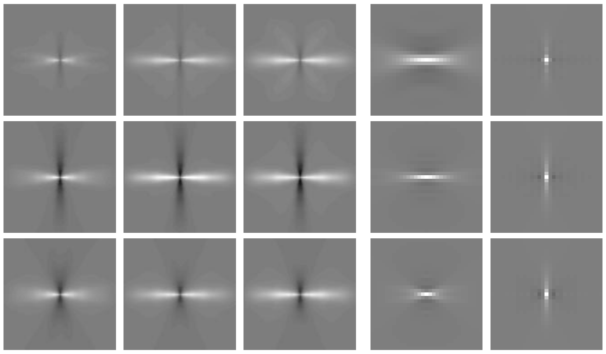

Figure 4. Spatial DST for Horizontal vs. Vertical artificial scenes.

If we compute the output of the two filters by convolution with the respective impulse responses $o_+(x,y) = i(x,y) * h_+(x,y)$ and $o_-(x,y) = i(x,y) * h_-(x,y)$, the structural semantic feature can be obtained as:

$$u = \iint |o_+(x,y)|^2 \, dx \, dy - \tag{10}$$
$$\iint |o_-(x,y)|^2 \, dx \, dy = E_+ - E_-$$

The impulse responses of these two filters are Receptive Fields that best discriminate between two groups of images and that allow a continuous organization of scenes. Fig. 1 shows the impulse responses of both filters for the three DSTs introduced in this paper.

DST (see Fig. 1) computed in the Fourier domain, is equivalent to the convolution of the image with two spatial invariant filters and computing the difference of their total output energies. The two impulse responses $h_+(x,y)$ and $h_-(x,y)$ reveal the spatial features that are discriminant between the two opposite sets of images. For artificial vs. natural scenes we see a cross impulse response vs. an isotropic (slightly oblique) impulse response (Fig. 1.a). For open vs. closed natural scenes we find an horizontal edge detector vs. an isotropic impulse response (Fig. 1.b). For Horizontal versus vertical artificial environments, the impulse responses are an horizontal vs. a vertical edge detector (Fig. 1.c).

Here, we ask about the relevance of using spatial variant filters despite of global filters. In fact, we can imagine that for some categories of scenes, the main structural components may vary in function of their position in the image. Such an operation may be critical for performing comparison task between similar scenes. For example, for open natural environments, we can expect to have an horizon in the middle center with texture at the bottom and sky at the top. However, closed environments may present texture everywhere in the image with sometimes, oblique shapes at the top. Spatial variant filters are expected to take these spatial differences into account for improving ordering performances along the axes.

To investigate this point, we divide images (256x256 pixels) in 9 overlapped windows (128x128 pixels) from left to right and from top to bottom. We compute the power spectrum of each window, $\Gamma_i(f_x, f_y)$, and we compute a structural feature which is a composition of the 9 DSTs and the 9 power spectra:

$$u = \sum_{i=1}^{9} \left(\iint \Gamma_i(f_x, f_y) \, DST_i(f_x, f_y) df_x \, df_y \right) \tag{11}$$

The 9 DST_i are obtained by the same learning procedure as for the global DST. Differences in the shapes of the 9 DSTs will reveal different statistics of the discriminant orientations and spatial frequencies. Figure 2, 3 and 4 show the 9 DSTs. It must be emphasized that both Global DST and Spatial DST compute eventually an unique semantic feature. We can look at the impulse responses of the two discriminant filters H_+ and H_- for each DST_i. The first observation is that the DSTs vary only from top to bottom whatever the semantic axis. This means that spatial variant filters depend only on the vertical spatial variable y and not on x. This is an expected result as both artificial and natural environments have a layered structure from top to bottom (main structures, object attributes and positions differ from top to bottom but not from left to right). Another interesting result is that when discriminating between artificial and natural scenes, the 9 DSTs are highly similar. In that case, the spatial arrangement of dominant orientations carries very low information for making the difference between artificial and natural scenes. A global measure of dominant orientations over the image gives enough structural information to resolve this categorization. On the contrary, computations of *degree of openness* (Fig. 3) and *degree of verticalness* (Fig. 4) is improved by spatially variant filters.

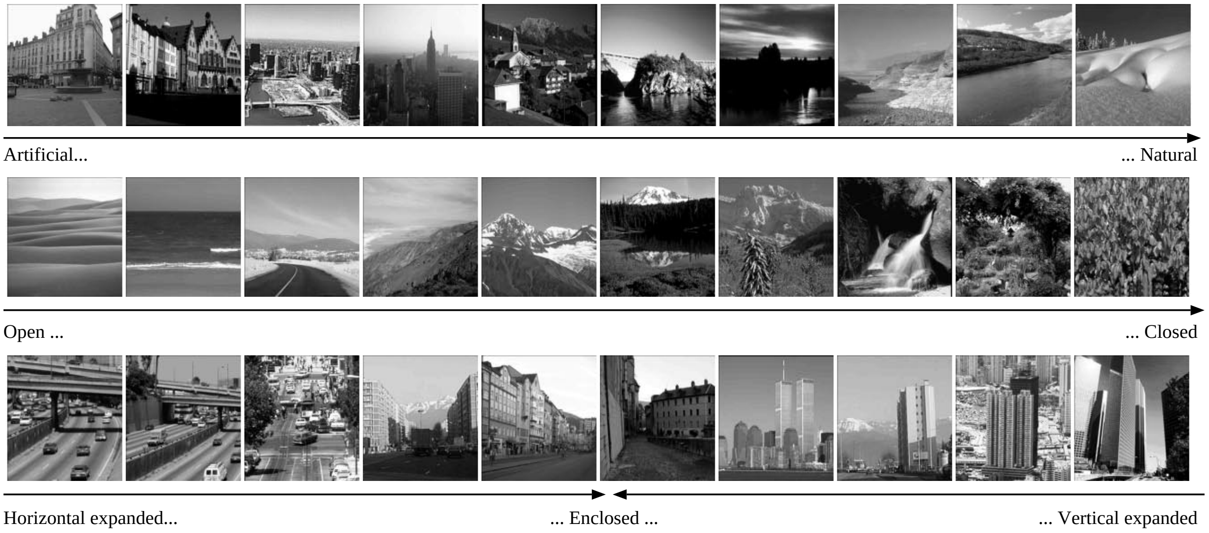

Figure 5. Organization of a sample of real-world scenes pictures along the semantic axes. From top to bottom: Artificial to natural scenes, open to closed natural scenes and horizontally to vertically expanded artificial scenes (at the middle of this axis are enclosed urban areas).

4 Semantic Ordering Procedure

In this section, we project scenes never learnt along the three semantic axes and provide elements of comparison between the Global DST and the spatially variant DST. We worked with gray-level pictures of 256x256 pixels size (from the Corel image database, Web pictures and personal pictures). The image database contains 2600 images (1500 natural scenes, 800 artificial scenes and 300 scenes containing both natural and artificial elements).

4.1 Ordering along the Artificial to Natural axis

The classification rate (obtained by cross-validation) for artificial and natural scenes is similar, using Global DST (92%) or Spatial DST (91%). The top line of Fig. 5 displays a sample of images revealing the organization along the Artificial to Natural axis (using the Spatial DST). Ambiguous images containing both man-made and natural structures are mainly located around the center of the axis.

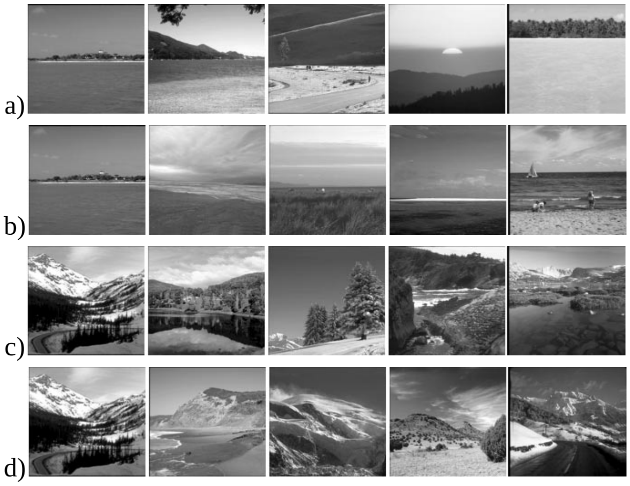

Figure 6. Similar natural scenes, using Global DST (a,c) and spatial DST (b,d).

4.2 Ordering along the Open to closed axis

This semantic axis organizes natural scenes from panoramic areas to closed and bounded natural environments. The classification rate is 94% with Global DST and 97% with Spatial Variant DST.

The middle line of Fig. 5 shows the continuous organization along the natural semantic axis (using 1500 natural scenes). We observe an interesting ordering: from open scenes with panoramic views (coast, beach, desert), progressively filled in with mountains (valley, mountain), until textural scenes (forest, waterfalls). Of course, some scenes categories may be mixed along the axis (coastline, desert and beach zones may overlap), but the addition of color information definitively disambiguates the semantic status of the scene [3].

[3] Performances using DST and color information is detailed in a manuscript in preparation by the two authors.

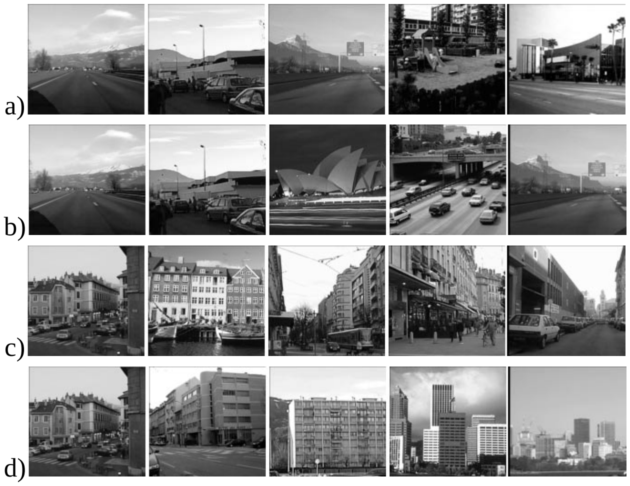

Figure 7. Similar artificial scenes, using Global DST (a,c) and spatial DST (b,d).

In order to compare performances of the two structural DST, we evaluated the ability to retrieve similar images to a given prototype. Results are shown on Fig. 6, using a beach and a valley as prototypes (left-hand side). When using the Global DST, the retrieved images have the same degree of openness but the boundary elements can change. When using the spatial variant DST, the retrieved images look slightly more similar.

4.3 Ordering along the horizontal to vertical axis

This axis organizes artificial scenes along the horizontal to vertical axis (Fig. 5, bottom). Both Global DST and spatial variant DST give the same performances (98%) when classifying prototypical urban images in horizontal- vs. vertical structured scenes. Three semantic zones emerge: highways scenes, center and city street zones, then city buildings and skyscrapers. Fig. 7 shows a sample of performances of both structural features in a retrieving task. Both results show very good performances knowing that images are retrieved by a unique structural feature (either Global DST, either Spatial DST).

5 Conclusion

In this paper, we introduced three semantic axes and computation of original filters (Discriminant Spectral Templates), optimum for the task at hand. Even if it is obvious that a few more axes would allow a precise classification, we observe that only two structural attributes, *degree of naturalness* and *degree of openness* of a scene, allow the emergence of semantic categories. Moreover, this procedure is very efficient for providing a low cost computational method, as once DSTs are built, computation of image coordinates along the axes needs only few operations. Finally, we observe that performances are almost equivalent for the global and the spatially variant filters, highlighting the relevance of a coarse and global encoding of the main structure of the scene for its recognition.

References

[1] M. M. Gorkani and R. W. Picard. Texture orientation for sorting photos ”at a glance”. Proc. Int Conf. Pat. Rec., Jerusalem, 1994, Vol I, pp. 459-464

[2] P. Lipson, E. Grimson, and P. Sinha. Configuration based scene classification and image indexing. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Puerto Rico, 1997, pp 1007-1013 (IEEE Computer Society Press)

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

[4] A. Oliva. Perception de Scenes [Scene Perception]. PhD dissertation, Institut National Polytechnique de Grenoble, France; May 1995.

[5] A. Oliva and P.G. Schyns. Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli. Cognitive Psychology 1997; 34:72-107

[6] A. Oliva and P.G. Schyns. Diagnostic color blobs mediate scene recognition. Cognitive Psychology, in press

[7] A. Oliva, A. Torralba, A. Guerin-Dugue and J. Heraut. Global semantic classification using power spectrum templates. The Challenge of Image Retrieval,(CIR99). Electronics Workshops in Computing series, Springer-Verlag. Newcastle, 1999

[8] B. D. Ripley. Pattern recognition and neural networks. Cambridge University Press, 1996

[9] D. L. Swets and J. J. Weng. Using discriminant eigenfeatures for image retrieval. IEEE Trans. On Pattern Analysis and Mach. Intell.1996; 18:831-836

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

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