Global Semantic Classification of Scenes using Power Spectrum Templates

Aude Oliva, Antonio B. Torralba
Anne Guerin-Dugue and Jeanny Herault
Laboratoire des Images et des Signaux (LIS)
Institut National Polytechnique de Grenoble (INPG)
Grenoble, France
{oliva,torralba}@lis-viallet.inpg.fr

Challenge of Image Retrieval (CIR99),
Elect. Work. in Computing Series, Sringer-Verlag, Newcastle, 1999.

Abstract

Scene recognition and content-based procedures are of great interest for image indexing applications processing very large databases. Knowing the category to which a scene belongs, a retrieval system may filter out images belonging to other categories.

In this paper, we introduce a computational approach which classifies and organises real-world scenes along broad semantic axes. Fundamental to our approach is the computation of global spectral templates providing a continuous organisation of scenes between two categories. These templates encode the structure which is discriminant between two categories. We propose a hierarchical procedure of two stages, that organises images along three semantic axis. Firstly, all the scenes are classified according to an *Artificial to Natural* axis. Then, natural scenes are organised along the *Open to Closed* axis whereas artificial environments are classified according to the *Expanded to Enclosed* scenes axis.

1 Introduction

Everyday complex scenes depicted in photographs and movies are recognised by the human visual system as rapidly as objects presented individually. Such an automatic and efficient recognition is currently a computational dream (or nightmare) for artificial visual systems. In particular, reaching such a level of performance is a critical feature of indexing of image databases.

Image retrieval systems usually represent images by a collection of low-level features such as colour, texture, edge positions and spatial relationships in the image. These features are used to compute the similarity between a picture selected by the user and the images in the database. The query is based on an image features vector matching between images while human classifications are based on fuzzy similarity computations that are often context-driven.

Bridging the gap between higher concepts such as “urban scenes” or “snowy mountains” and low-level features extracted from the picture, requires two fundamental operations: finding the relevant semantic description of the concept and finding the relevant low-level features.

In this paper, we present the first results of a computational approach that classifies and organises real-world scenes along semantic axes. The research introduces the main concepts of the approach and describes classification results in ”super-ordinate semantic classes” [1]. The main marks of our approach are two-fold:

[1] A *super-ordinate* category (e.g. artificial or natural scenes, urban areas, horizon landscapes, indoor scenes, etc.) can be

1) determine the relevant spectral features correlated with the semantic content of the image 2) a *continuous organisation* of scenes.

The low-level features set is represented by a template. This template uses the spectral content of the image (i.e. the energy distribution through spatial frequencies and orientations) in order to discriminate between images with different semantic contents. We propose to call it a *Discriminant Spectral Template* (DST). The main properties of a DST are the following: 1) it is dependant upon a specific semantic image contain 2) it allows to *continuously* organise pictures along a specific semantic axis and 3) it is robust to image variability.

In the following, we describe the organisation of real-world scenes along three semantic axes, each one providing an ideal DST (e.g. from artificial scenes to natural scenes; from “open” landscapes to “closed” landscapes, and from ”city scenes” to “enclosed” urban scenes and indoors).

2 Context-Driven Recognition

Context-driven recognition procedures usually assume that a semantic classification can emerge from very simple computations based on low-level features [3-5,8,11,16,17]. Knowing the meaning of the scene, a retrieval system may compute in advance its semantic category.

Recent studies attempt to address this complex issue. For instance, Lipson et al. [6] reasoned that scene categories should be invariant to image transformations such as scaling, illumination and precise objects location. They encode the global configuration of the scene by using spatial and photometric relationships within and across regions of images. Even though it is effective for scene categories that are geometrically well-defined (e.g. snowy mountains with blue sky), their method cannot be generalised to broader categories or scenes where parts and objects are randomly localised (such as rooms or indoor buildings). In a similar vein, Picard and collaborators [3, 7, 16] represent scenes by a collection of features (texture, colour, spatial frequencies) locally computed on tessellated images. Their strategy can retrieve images of urban scenes from landscapes [3] or classify pictures in indoor vs. outdoor categories [16]. The work of Vailaya et al. evaluates the discrimination power of several low-level features in order to classify city images versus landscapes [17]. Common to all these approaches is the classification into exclusive classes. However, when dealing with large databases, exclusive classification may increase irrelevant classification rate as most of pictures are ambiguous in terms of category.

Fundamental to our approach is the notion that scene recognition requires the definition of *continuous* semantic axis. The keystone is the computation of Discriminant Spectral Templates, encoding the spectral features that better discriminate between two categories.

3 Power Spectrum Families

The search for an unique template came from experimental results about how the human visual system may recognise a complex scene. Typically, real-world scenes belonging to the same category tend to have a similar organisation of their main component objects. Oriented shapes of the main components define the “skeleton” of the spatial organisation [14]. Human visual processing seems to use these spatial regularities to categorise a picture in broad classes (e.g. urban areas, coastlines, landscapes, rooms, textured environments, ...) [9,10,14]. This “express” visual categorisation is based on a coarse invariant information which is independent of the viewpoint, object locations, occlusions, shadows, colour and illuminant variations.

To this extent, the power spectrum turns out to be a very good candidate for encoding such a structural information. The power spectrum of an image is the square of the magnitude of its Fourier transform. It gives global information about the basic elements that form the image. Here, we are not interested in a detailed analysis of the power spectrum which would be as complicated as studying the pixelised image itself. We only look for global characteristics in terms of the main orientations in the image, the dominant spatial

described as a broad semantic category which subsumes several *basic-level* scene categories (respectively cities or mountains, city center or streets, panoramic beaches or valleys, kitchens or bedrooms).

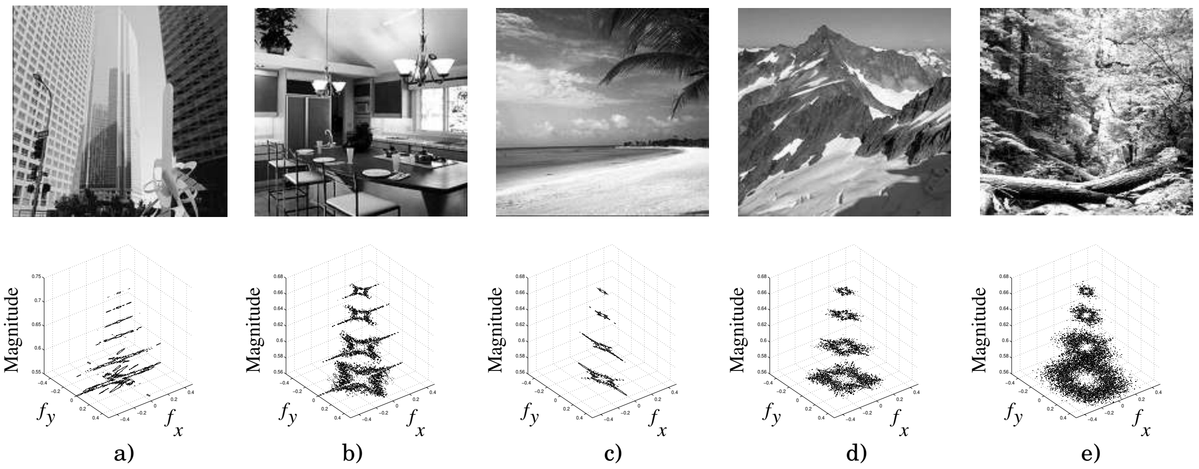

Figure 1: Examples of power spectrum forms for prototypical images (vertical axis is the magnitude in logarithmic scale, horizontal axis are the spatial frequencies f_x and f_y). At the bottom, we show sections at several levels of the power spectrum of each image.

scales, the periodic patterns, etc. Power spectrum answers to questions such as: which frequency band encodes most of the energy; what is its global form through spatial scales; are there poles and narrow lines of energy? All these features would mostly be independent of objects arrangements, point of view, illumination, etc. For example, typical beach scenes have a strong horizontal organisation (Fig. 1 c). Therefore, their power spectra will display a dominance of energy on the f_y axis, mostly at low f_y spatial frequencies. At the opposite extreme, city scenes are usually structured along vertical and horizontal directions whereas spectra of forests would be mainly isotropic from low to high spatial frequencies (some examples are shown Fig. 1).

The power spectra of real-world images exhibit very different energy distributions for each orientations and spatial frequencies. In analysing images from a wide set of real-world environments, we observed a strong bias towards horizontal and vertical orientations [1, 2, 13].

We observed that five families of power spectrum can be defined showing a strong correlation with the semantic content of the image [18]. Figure 1 shows prototypical power spectra for the five proposed families. These families are characterised by the shape of their dominant orientations:

1. *Horizontal shape*: The power spectrum exhibits an horizontal dominant line, on the f_x axis, from low to high spatial frequencies. A good example is a city scene composed of tall buildings (Fig.1a).

2. *Cross shape*: Vertical and horizontal directions are represented approximately equally in the power spectrum. A typical scene exhibiting such a cross form at all spatial frequencies is an indoor scene of a kitchen or a living-room, mainly composed of man-made objects of small and medium sizes (Fig.1b).

3. *Vertical shape*: The power spectrum shows a vertically dominant line (f_y axis) revealing that the scene has an horizontal structure. Examples are beach and field scenes, as well as other panoramic scenes (Fig.1c).

4. *Oblique shape*: Oblique orientations (mainly orientations at 45 deg plus or minus 15 deg) dominate the power spectrum. Examples are images of mountain areas, canyons, valleys (Fig.1d).

5. *Circular shape*: All the orientations are equally represented in the picture, leading to an isotropic power spectrum. Common examples are highly textured environments such as forests, fields (Fig.1e).

Fig 1 shows typical examples of scenes belonging to each of the five power spectrum categories outlined above. For these prototypical images, the shape of the power spectrum is conserved across spatial scale.

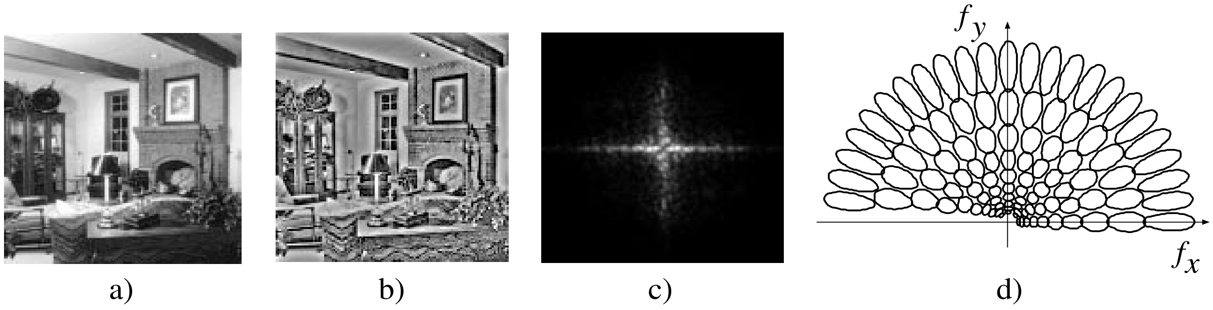

Figure 2: This figure shows the main steps for computing the vector of 100 components used to represent an image. a) Original image. b) Output of the pre-processing stage. The effect of illuminant and shadows have been reduced. c) Power spectrum of the prefiltered image. It is computed as the squared of the magnitude of the Fourier Transform. d) -3dB sections of the set of gabor filters used to sample the power spectrum. The highest frequency is 1/3 cycles/image and the lowest one is 1/72 cycles/image.

For most images however, the shape of the power spectrum varies gradually from one of these categories to another. For example, a field scene may display a vertical dominance at low-spatial frequencies corresponding to the horizon, and a "ring" at medium and high spatial frequencies corresponding respectively to the texture of the trees behind and the texture of the grass in front. As a consequence, the variety of natural images and their "intertwined" distributions of orientations is at the origin of the continuity along the semantic axes defined in this paper.

Instead of searching for low-level features (e.g. red and yellow) describing a specific semantic category (e.g. sunset beaches), we looked for semantic categories that would naturally emerge from the five major power spectrum forms. From the five main power spectrum forms displayed on Figure 1, we propose a hierarchical classification procedure as follows: a first level of classification discriminates between artificial vs. natural environements. The Horizontal and the Cross shapes together represent artificial environements (e.g. man-made scenes), whereas the three other shapes are typical of natural scenes [2]. Following this initial classification, the second level assesses natural scenes along an axis representing scenes from Open to Closed environments (e.g. open scenes are mainly horizontally structured with depth view –beaches, fields– whereas closed scenes are bounded environments, highly textured –forests, mountains. Open environments have a vertical spectrum shape and closed environments have circular and oblique spectrum shapes). This second level represents also artificial scenes along an axis revealing the vertical dominant structure of man-made outdoor and indoor environments that we call Expanded-Enclosed axis. This axis represents a continuum between unbroken areas of urban scenes made with tall and large buildings (horizontal spectrum shape) and confined images of indoor buildings and rooms (cross spectrum shape).

4 Computational Model

4.1 Image database

We chose 700 pictures from the Corel Image database so as to cover a large variety of real-world scenes. We imposed the constraint that images must not be pictures of isolated objects. Examples of scenes included beaches, fields, forests, mountain areas, deserts, waterfalls, canyons, urban areas such as shopping centers, streets, highways, skyscrapers and different kind of rooms. Out of the 700 images, 300 were classified as artificial environments (man-made scenes), another 300 were classified as natural, and the remaining 100

[2] As outlined in the previous section, classification must be done on continuus axes. Thus, a panoramic view with an urban area in the background would be located between the artificial and the natural poles, exhibing a power spectrum form having both a cross form and a vertical form.

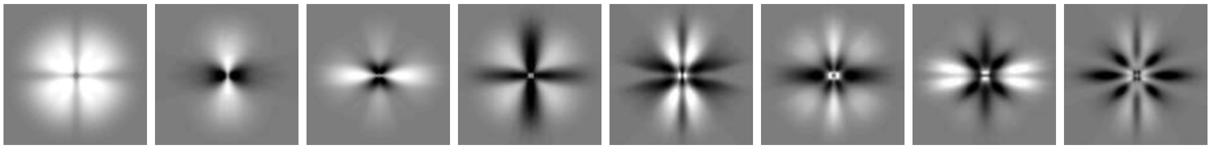

Figure 3: The first 8 Principal Components calculated from the power spectrum of 700 scenes. The horizontal coordinate is f_x and the vertical one is f_y. The symmetrical structure of the principal components is due to the mirror transformation applied to the image power spectrum.

were ambiguous scenes, namely natural scenes containing man-made objects (e.g. farm buildings in a field, benches in a garden, boats in an harbour, bridges over a ravine, ...). This classification was obtained by asking 4 observers to place each image in the artificial or the natural group. Images were 256 by 256 pixels in size, coded in 8-bit grey-levels.

4.2 Pre-processing

The aim of pre-processing is two-fold: reducing the effects of large shadows that may hide important parts of the scene and minimising the impact of high contrasted objects which would disturb the power spectrum shape of the background image. Firstly, we apply a logarithmic function to the intensity distribution. Then, we attenuate the very low spatial frequencies by applying a high pass filter. We apply an adjustment of the local standard deviation at each pixel of the image. This operation makes large regions of the image being equally bright (see Fig 2b).

4.3 Global Semantic Axes

The aim of the approach is to determine a unique template that can be applied to the power spectrum of an image so as to localise the image along a continuous one-dimensional semantic axis. We compute three different templates (DST) corresponding to the three semantic axes of the hierarchical procedure (artificial to natural scenes, open to closed scenes for natural environments and expanded to enclosed scenes for artificial environments).

The computation phases are as follows:

1) After the pre-processing stage, we compute the power spectrum (see Fig 2c) and we sample it with a set of narrow band Gabor filters (see Fig. 2d).

2) We use Discriminant Analysis in order to compute the axes. To define each axis, we have chosen two sets of prototypical scenes in order to set up the extremities of the axis. The Discriminant Analysis computes the axis that both maximises the distance between the two prototypical groups and minimises the standard deviation of the images belonging to the same group.

4.3.1 Spectral Representation

If we compute the power spectrum of an image of size 256x256 pixels (by computing the magnitude of the Discrete Fourier Transform of the image), we obtain a 256x128 (discarding the radial symmetry of the power spectrum) vector of low-level features for each image. Therefore, images are distributed in a very high dimensional space. To reduce dimensionality, we sample the power spectrum by a set of narrow-band Gabor filters (100, see Fig. 2d) tuned to different spatial frequencies (orientations and scales) from low spatial frequencies (1/72 cycles/image) to high spatial frequencies (1/3 cycles/image). We sample the power spectrum with 10 spatial frequency radial bands and a decreasing number of orientations from high (24) to medium (12) and low (4) spatial scales.

The transfer function of a Gabor filter tuned to the spatial frequency f_r in the direction determined by the angle θ is given by the expression:

$$G(f_x, f_y) = K\, e^{-2\pi^2\left(\sigma_x^2 (f_x' - f_r)^2 + \sigma_y^2 f_y'^2\right)} \tag{1}$$

where f_x' and f_y' are obtained by rotation of the spatial frequencies $f_x' = f_x \cos(\theta) + f_y \sin(\theta)$ and $f_y' = -f_x \sin(\theta) + f_y \cos(\theta)$. σ_x and σ_y give the shape and frequency resolution of the Gabor filter. K is a constant. The full set of filters is obtained by rotation and scaling of this expression. This gives a high frequency resolution at low spatial frequencies and a low frequency resolution at high spatial frequencies. The values σ_x and σ_y are chosen in order to have coincidence in the contour section of the magnitude at -3dB.

Given an image, its semantic content is invariant with respect to an horizontal mirror transformation of the image. Therefore, we compute the symmetric energy outputs of the Gabor filters which are invariant with respect to an horizontal mirror transformation:

$$\Gamma_{f_r,\theta} = \int\int |I(f_x, f_y)|^2 \left[G^2_{f_r,\theta}(f_x, f_y) + G^2_{f_r,\pi-\theta}(f_x, f_y)\right] df_x\, df_y \tag{2}$$

where $|I(f_x, f_y)|^2$ is the power spectrum of the image. $G_{f_r,\theta}$ and $G_{f_r,\pi-\theta}$ are two Gabor filters tuned to the spatial frequencies given by the radial frequency f_r and the directions θ and $\pi - \theta$. Therefore, the value $\Gamma_{f_r,\theta}$ is invariant with respect to an horizontal mirror transformation of the image.

The features we are going to use are the normalised ones:

$$\widetilde{\Gamma}_{f_r,\theta} = \frac{\Gamma_{f_r,\theta} - E\left(\Gamma_{f_r,\theta}\right)}{std\left(\Gamma_{f_r,\theta}\right)} \tag{3}$$

where E and *std* are the mean and the standard deviation of the features $\Gamma_{f_r,\theta}$ computed over the entire image database. Therefore, for each image, we have a feature vector defined by the collection of $\widetilde{\Gamma}_{f_r,\theta}$ obtained at different frequencies and orientations.

4.3.2 Discriminant Spectral Template

Before applying the discriminant analysis, we performed a second dimensionality reduction (from 100 to 8). We computed the principal components (PC) of the normalised energy features (eq. 3) over the entire image database. Principal Component Analysis (PCA) gives the orthogonal axes (called principal components) that best represent the variance of the distribution. This operation reduces the dimensionality by taking into account only the most important components, the components that are responsible of the variability between images in the feature space [15]. Figure 3 shows the eight first Principal Components computed from the entire database. The PC pictures are computed by addition of the set of Gabor filters, weighted by the components of the principal vectors obtained by the PCA. The symmetrical structure of the principal components is due to the mirror transformation originally applied to the features [3] (eq. 3).

After projection of each image spectral features onto the principal components, we compute the semantic axis by applying Discriminant Analysis. The following method is applied to each of the three semantic axes. Firstly, we separate the database in two groups defining the extremities of the axis. For this purpose, we only use prototypical images. Discriminant Analysis consists in the search of the axis maximising the distance between classes while minimising the dispersion between elements of the same class [see [12] for a descriptive review of the method]. Half of the image database is used for the learning stage and the other half for the testing stage.

As two prototypical groups of images are considered, Discriminant Analysis provides only one discriminant vector. This discriminant vector is expected to represent a relevant organisation of pictures along the semantic axis we are looking for. Therefore, we compute the projection of each image spectral features on

[3] Note that computing the PCA directly from the Gabor outputs without the mirror transformation shows poorer performances, because only some of the principal components are nearly symmetric.

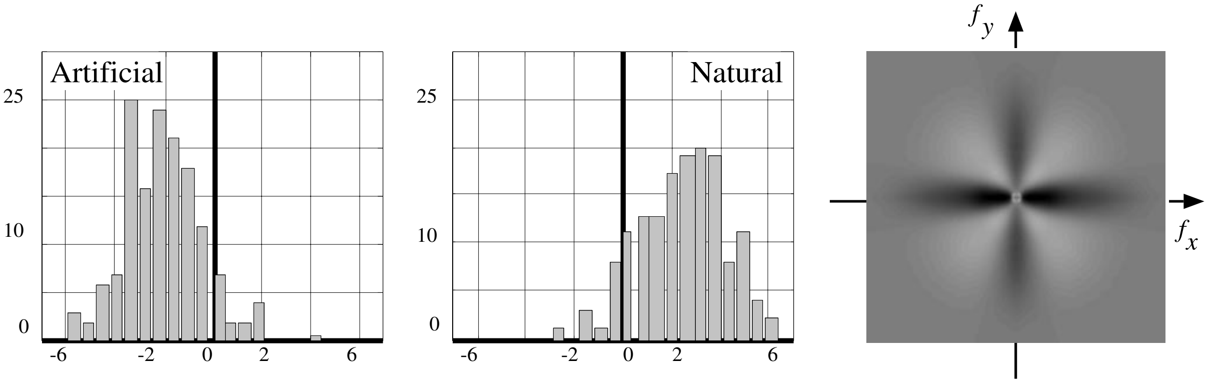

Figure 4: Results of projection of the testing group onto the Artificial-Natural axis. The histograms show the distribution of both the artificial and the natural images sets onto the axis (testing phase). 90 % of the images are accurately classified in the testing phase (91 % in the learning phase). At the righthand side, we show the resulting DST. The angular anisotropy reveals the importance of the statistics of dominant orientations between different groups of images.

the semantic axis. Figures 4, 7 and 9 display the three discriminant vectors. As in Figure 3, these pictures are obtained by addition of the set of Gabor filters, each one weighted by the components of the discriminant vector. They have been computed from the eight first principal components. We call this representation a *Discriminant Spectral Template* or DST.

4.4 Artificial-Natural Axis

The objective is to compute the DST associated with the Artificial-Natural axis. Artificial scenes are composed of man-made objects, having dominant vertical and horizontal edges. Thus, their power spectrum displays an horizontal shape or a cross form (see Fig. 1 a and b). On the contrary, the orientation distribution in the power spectrum of a natural scene is usually isotropic (Fig. 1 e) or has an oblique dominance (Fig. 1 d). Other natural scenes have a strong horizontal component due to the horizon (Fig. 1 c). Out of 600 artificial and natural images, 300 prototypical scenes have been used in the learning phase (150 for each class) in order to compute the Artificial-Natural DST, as described in the previous section.

The validity of the DST is assessed along two criteria. We desire that the classification rate obtained by the discriminant analysis approximates the classification obtained by humans in the two-alternative forced choice task (Artificial vs. Natural). Then, the pictures should organise themselves in a coherent and dense way along the considered axis.

In the testing phase, 300 new pictures are projected using the DST shown in Figure 4. The DST shows how the spectral components should be weighted in order to differentiate artificial from natural environments: the white and dark parts respectively represent natural and artificial components. Natural components are found at very low vertically oriented spatial frequencies and oblique orientations at all spatial scales. Artificial components describe a cross, strengthened along the horizontal at low and medium spatial frequencies. Results of the classification in exclusive groups are displayed Figure 4 (the histograms). Both learning and testing performances are slightly larger than 90 %.

The second criterion has been tested in two ways. First, Figure 5 displays prototypical pictures randomly selected and equally spaced along the axis. We observed that the left side exhibits artificial scenes and the right side exhibits natural scenes.

To fully investigate the relevance of the artificial-natural DST, we projected 100 other new pictures, supposed to be "ambiguous" regarding their artificial or natural status (e.g. natural environments containing more or less man-made structures).

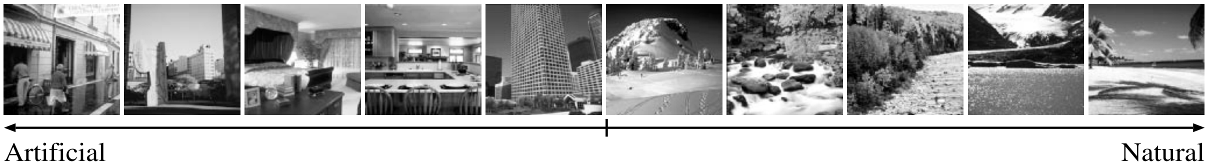

Figure 5: Organisation of new prototypical scenes on the Artificial-Natural axis. Pictures have been randomly selected and are equally spaced along the axis. The left side exhibits artificial scenes and the right side exhibits natural scenes.

Figure 6 shows a sample of the results: the images with an underlined segment are prototypical pictures displayed to show the extremities of the axis. Interestingly, the new ambiguous images are mainly projected around the middle of the axis (represented by the center line of Figure 6) rather than the extremities. The bottom line displays pictures seen as mainly natural: indeed, the farm scene contains a dominant textured field and the panoramic view over the village scene is strongly horizontally structured like a natural open landscape. Thus, the model prefers to classify these pictures as “natural”. In a similar vein, the top line shows four ambiguous pictures considered as “more” artificial. In fact, their power spectrum exhibits a cross form style. The medium line of Figure 6 represents pictures classified around the center of the axis. Note that the right images in the central line, are horizontally structured: at low spatial frequencies, vertical components (f_y) are dominant, corresponding to the natural components of the DST (white). Note also that the images left of the medium line are mainly composed of artificial objects (the boat and the farm). Their power spectra are closer to the artificial components of the DST (dark).

4.5 Open-Closed Axis

Knowing that a picture belongs more to the natural or the artificial set, we propose to compute another DST which purpose is to organise natural scenes along a specific semantic axis (*Open to Closed* axis). This axis originally comes from three power spectrum shapes (cf. Fig. 1-c-d-e). Looking at natural environments, we propose to represent scenes from open, unbroken areas with an horizon (*Vertical* Power Spectrum family) to bounded and closed textural environments (*Circular* and *Oblique* Power Spectrum families). Ideally, imagine a scene with a prominent horizon (e.g. a panoramic view on a valley) being slowly “filled in”, either by distant large and tall objects in the background (e.g. mountains) or by closer textured parts (e.g. trees, bushes, grass).

From the 300 prototypical natural images (e.g. beaches, seashores, oceans, deserts, fields, various landscapes, forests, gardens, waterfall areas, snowy or rocky mountains, valleys, open and closed canyons, ...), 150 were used for the learning phase and the DST computation and 150 for the testing phase. Learning and testing performances of classification rate in Open-Closed are about 88 % (cf. histograms of Figure 7). [4]

Figure 8 displays exemplars of natural scenes randomly selected and equally spaced along the axis. We observe an appropriate organisation from open areas (left side of Figure 8) such as fields, coastlines, panoramic valley views, progressively replaced by mountains environments and wide textured scenes (e.g. gardens, close-up bushes views, forests,...). The Open-Closed DST is displayed in Figure 7. The dark parts (negative values) represent the open components whereas the white parts (positive values) represent the closed ones. The intensity of dark and white parts reveals how the spectral components of a natural image should be weighted to compute its position on the axis.

[4] Note that the rate in exclusive classification is only indicative of the projection of prototypical images. The purpose of the DST representation is to look for a continuous organisation from Open to Closed environments.

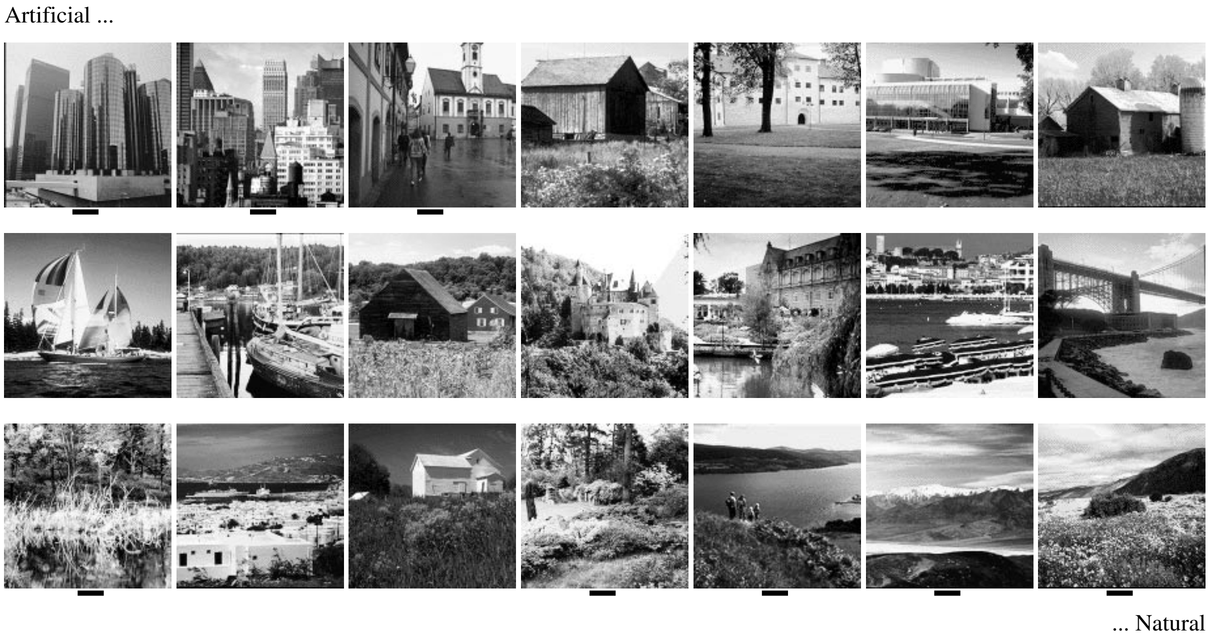

Figure 6: Examples of ambiguous scenes and their organisation along the Artificial-Natural axis. Images are sorted according to the Artificial-Natural DST: from the top to the bottom and from the left to the right, scenes are organised from the most artificial to the most natural. Underlined images belong to the prototypical groups.

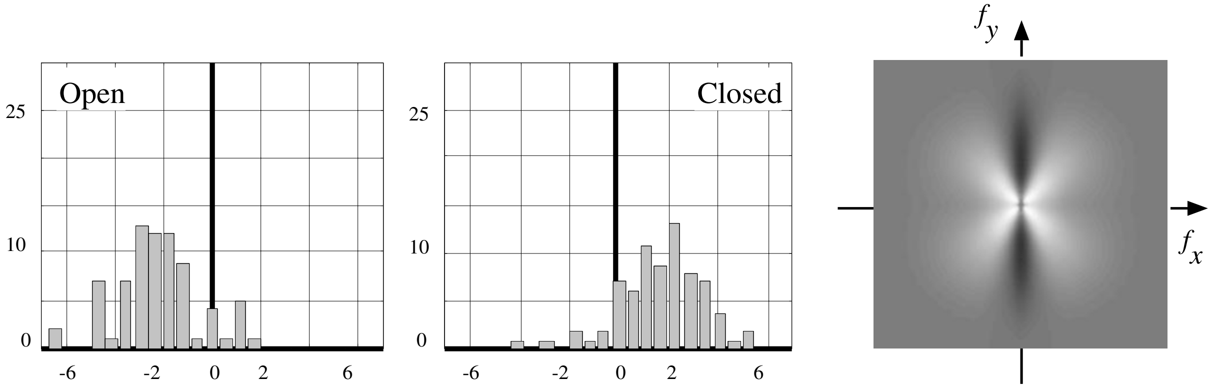

Figure 7: Results of projection of the testing group of natural images onto the Open-Closed axis. At the rigthhand side we show the resulting DST. 88 % of images are well-classified both in the learning and the testing phases.

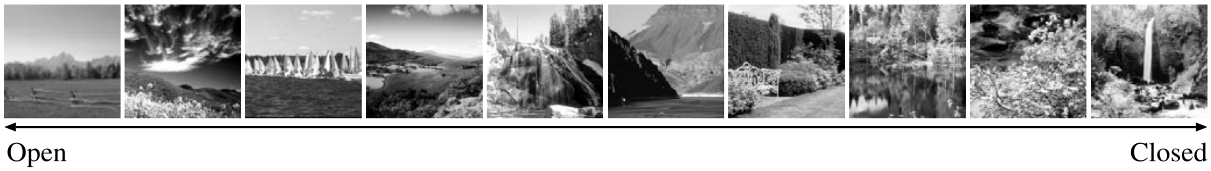

Figure 8: Organisation of new scenes randomly selected from the testing group along the Open-Closed axis.

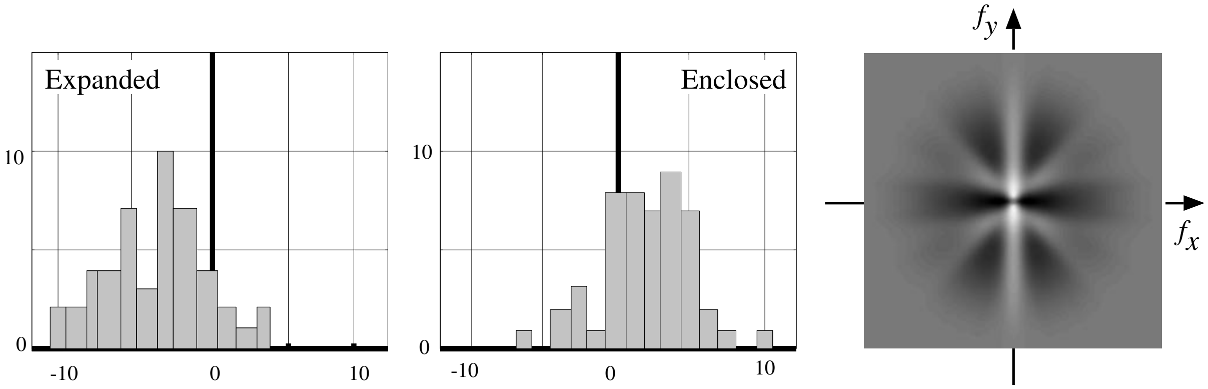

Figure 9: Results of projection of the testing group of prototypical artificial images onto the Expanded-Enclosed axis. At the rigthhand we show the resulting DST. 82 % of images are well-classified.

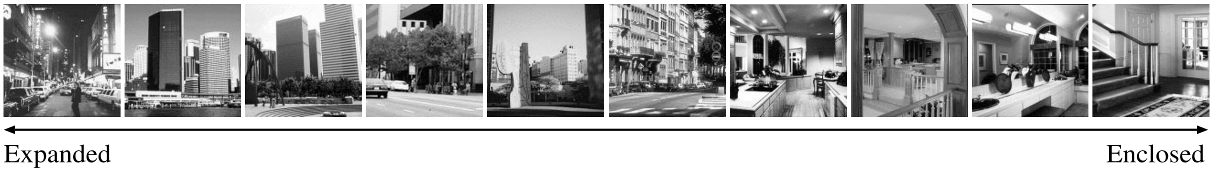

Figure 10: Organisation of new scenes randomly selected from the testing group along the Expanded-Enclosed axis.

4.6 Expanded-Enclosed Axis

In a way similar to the analysis along the Open-Closed axis, we looked at a possible super-ordinate semantic for the artificial scenes. From the *Horizontal* and the *Cross* power spectra shapes (Figs. 1 a and 1 b), two broad categories seem to emerge: *Expanded* areas of urban scenes with vertically structured parts and confined and *Enclosed* scenes as urban areas and indoors. The image database was composed of 200 artificial scenes representing prototypical images grouped either in the Expanded or Enclosed classes. The learning stage was performed on half of the database (100) and used to define the Expanded-Enclosed DST (see Figure 9).

The classification rate both for learning and testing was 82 % (Figure 9). Note that this two-alternative forced choice task (Expanded vs. Enclosed) is ambiguous for human subjects. The organisation is shown in Figure 10. A careful look at the selected pictures, from left to right, shows a “broad” regularity of the organisation chosen by the model, e.g. from vertically structured areas, some with strong vanishing lines (see the left picture of the Figure 10), to indoor scenes characterised by a double dominance of horizontal and vertical edges. The dark parts of the Expanded-Enclosed DST represent the Expanded components and the white parts represent the Enclosed components. The dark Expanded components are located along the horizontal direction (f_x) and at an orientation plus or minus 30 deg. around (f_y). These latter components may characterise the vanishing lines of images with perspective views. The white Enclosed components are vertically displayed.

5 Conclusion

In this paper, we present a novel computational method for performing broad semantic categorisation. Images are organised along semantic axes using a hierarchical representation (artificial to natural scenes, open to closed natural scenes and expanded to enclosed artificial scenes). The position of an image along each axis is obtained by matching its power spectrum with a *Discriminant Spectral Template* (DST) computed from learning over a subset of the image database. The DSTs appear to be a suitable representation for *continuously* organising pictures along semantic axes and thus, taking into account the ambiguous nature of real-world scenes.

Scene recognition algorithms and context-driven procedures are of great interest for image indexing applications processing very large databases. Our approach offers a collection of features (DST) [5] that appears to be strongly correlated with a meaningful content of the scene.

Acknowledgments

The authors wish to thank two anonymous reviewers for helful comments and Dr. P. Mamassian and K. Davies for discussions of an earlier version of this manuscript. This work was funded by a post-doctoral fellowship awarded to Aude Oliva from the *Region Rhones-Alpes*, ACTIV project on Image Indexing, and by a PhD. fellowship from the *MEESR*, France, awarded to Antonio B. Torralba. Please, send all correspondance concerning this paper to A. Oliva or A.B. Torralba {oliva, torralba}@lis-viallet.inpg.fr.

References

[1] R. Baddeley. The correlational structure of natural images and the calibration of spatial representations. Cognitive Science 1997; 21:351–372

[2] D. J. Field and N. Brady. Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes. Vision Research 1997; 37:3367–3383

[3] 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

[4] A. Gurin-Dugu and A. Oliva. Natural image classification from distribution of local dominant orientations. 11th Scandinavian Conference on Image Analysis. Kangerlussuaq, Greenland, 1999

[5] J. Hrault, A. Oliva, and A. Gurin-Dugu. Scene categorisation by curvilinear component analysis of low frequency spectra. In European Symposium on Artificial Neural Networks. Bruges, 1997, pp 91-96

[6] 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)

[7] 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

[8] A. Oliva. Perception de Scnes [Scene Perception]. PhD thesis, Institut National Polytechnique de Grenoble, 1995

[9] A. Oliva and P. G. Schyns. Color influences fast scene categorization. In Proceedings of the 18th annual conference of the cognitive science society. San Diego, California, 1996, pp 239-242

[10] 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

[11] A. Oliva and A.B. Torralba. Scene Semantic Spaces from Global and Color Templates. Technical Report, LIS-INPG, March 99.

[5] The DSTs proposed in this paper are not exhaustive and other DSTs may define other semantical axes.

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

[13] A. Schaaf and J. H. Hateren. Modeling the power spectra of natural: statistics and information. Vision Research 1996; 36:2759-2770

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

[15] D. L. Swets and J. J. Weng. Using discriminant eigenfeatures for image retrieval. IEEE transactions on Pattern Analysis and Machine Intelligence 1996; 18:831-836

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

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

[18] A. B. Torralba and A. Oliva. Semantic Organisation of Scenes using Discriminant Structural Templates. IEEE International Conference in Computer Vision (ICCV'99)