ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 91-96

Scene Categorisation by Curvilinear Component Analysis of Low Frequency Spectra

Jeanny Hérault[1,3], Aude Oliva[2], Anne Guérin-Dugué[1]

[1] INPG-TIRF, 46 Avenue Félix Viallet, 38031 Grenoble Cedex, France
[2] Dept. of Psychology University of Glasgow, Glasgow, UK
[3] ISTG-UJF, Grenoble. (herault@tirf.inpg.fr, aude@psy.gla.ac.uk)

Abstract. This paper presents an original approach to scene categorisation by image analysis on a perceptual point of view. For computer vision, it is a very complex task to extract relevant information from real scene images (city, forest, room...) to categorise them. On the contrary, human visual system has extraordinary capacities to do this very quickly. We propose here a categorisation strategy based on low and very low spatial frequency power spectra and show that this information is sufficient to produce homogeneous scenes clusters. A new self-organising neural network (CCA) realises a non-linear projection into a bidimensionnal output space able to enhance the cluster structure.

1. Introduction

Complex image recognition is becoming one of the most challenging problems in computational vision with the explosive growth of visual databases [8]. These databases consist of thousands of images, used in an increasing number of applications, such as criminal identification, multimedia encyclopaedias, art and trademarks description. Image indexing usually requires human intervention to label every image coming into the database and to characterise its components.

Most of the images are composed of several objects. This object information is not the only relevant information needed to correctly label the image meaning. A scene characterised by the keywords *car* and *truck* could be a highway scene, a car park, a traffic jam, or a set of toys. Others cues about the image (i.e. the spatial relationships between objects) need to be used. Therefore, automatic indexing of the image *category* should greatly facilitate the image recognition procedure. This category information provides information about the context in which the objects may be found

There is now a great deal of evidence that the visual system identifies the category to which the scene belongs before identifying the objects [2,6,12]. Moreover, scene recognition can be initiated by coarse information available at low resolution, without using any information about the objects' identity [12]. This *global* information is automatically processed at the beginning of the visual processing. Thus, the object search strategy should be greatly improved by a *context-directed* strategy: It is well established that an object is recognised faster and with a higher accuracy in congruent contexts (i.e. a car in a city, a chair in a room) than in incongruent contexts [6]. This observation emphasises that the best strategy for object recognition should be to first activate its *context.* This provides constraints on the later object recognition stage. Unfortunately, this strategy, although known to be used by the human visual system, has not received the attention it deserves in the computational vision.

2- Image representation

Real world images contain statistical regularities that set them apart from random or textural images. A number of studies have shown that the power spectrum profile of natural images depends as $1/f^2$ on the spatial frequency [5,14]. Due to the spatial frequency distribution of natural images, these authors observed that the average spectral power is anisotropically distributed over orientation (advantage for 0° and 90°). However, the spatial orientation of cells depends strongly on the visual environment distribution. More precisely, neurophysiological studies observed that the distribution of sensors over orientations favours 0° and 90° for high spatial frequencies, whereas the sensor number is equivalent, whatever the orientation for low spatial scale coding [3]. This differential cell density distribution according to spatial resolutions indicates that the visual system has adapted in order to maximise the visual input variations coding low spatial resolutions. Therefore, we predict that the low spatial frequency sensors are more applicable to discriminate between different visual inputs.

However psychophysical and neurophysiological studies did not take into account the differences between visual inputs, i.e. the image category. A closer look at the amplitude spectrum of different images shows that the distribution over orientation and spatial frequency varied widely from one scene category to another. Images belonging to the same category (i.e. visual *context*)have a similar global structure, whereas different categories seem to have a specific organisation. For example, city scenes are structured along horizontal and vertical axes, whereas beaches are mostly composed along the horizontal axis and forest scenes details are coded in the high spatial frequency part of the spectrum. Whatever the objects' position in the scene, the global structure defining the scene category is quite invariant. A local measure which takes into account the precise localisation of object contours might not be the most relevant measure for context recognition. The global computation is the one preferentially made by the human visual system for recognition at a glance and allows scene context interpretation without memorising the number, the localisation and the meaning of the objects belonging to the scene [9].

So far, relative spatial relationships between objects should define the most relevant property of a particular scene, thus allowing the observer to automatically and rapidly identify the scene structure. Schyns & Oliva [12] have tested this hypothesis and have proposed that scene categories (i.e..as city, road, room, forest, mountain, etc...) might have a typical and distinct spatial organisation of their major components, this structural information being encoded in resolution as low as 2 cycles/deg. (or at 1/32 of the sampling frequency, for a 256x256 image which subtends 4 degrees of visual angle). These spatial regularities are preferentially used by the visual system in rapid picture classification tasks, and allow our visual cognition to recognise the *context*, before the object identification stage occurs.

In computational vision, artificial systems have been unable to imitate the capacity of the visual system to pick up relevant information despite the numerous edges, boundaries, shadows, textures and colours. Scene recognition is considered to be the most complex recognition process: classical views argue that a scene is composed of objects, themselves composed of smaller units. From this hierarchical view, artificial vision deduces that scene understanding could be achieved by *a local-to-global* strategy. This strategy segments the picture into smaller parts, (from regions to objects to contours), in order to classify them, before rebuilding the whole scene context. The powerful human analysis is incompatible with such a time consuming strategy. Then, the scene recognition level which seems to be the first recognition step of the human visual system becomes the final one of artificial image analysis systems.

In this paper, we describe a new strategy for an artificial visual system engaged in complex visual information processing, taking into account some basic aspects highlighted in recent psychological studies.

3- Spatio-temporal preprocessing of images in the retina

Psychophysical studies have now established that the visual input is filtered at several spatio-temporal frequencies. By using a new methodology, *hybrid stimuli*, S-O have shown that, for rapid scene classification, the low resolution information in an image is preferentially used by the cognitive system; fine and precise information about object identities is neglected. Such a *coarse-to-fine* analysis is naturally computed by low level visual modules. It has been shown in a model of the vertebrate retina [1, 7] that time and space variables are not separable in the retinal outer plexiform layer. This leads to very interesting behaviour: the spatial transfer function of this early visual stage evolves in a *coarse-to-fine* manner when stimulated suddenly with a new image.

During the first 15-20 ms, the spatial transfer function is of a rather broadband low-pass type, and later becomes high-pass (Fig.1, left). When the stimulation is removed, the high-spatial frequencies are rapidly cancelled and a very low spatial frequency rebound appears (Fig.1, right).

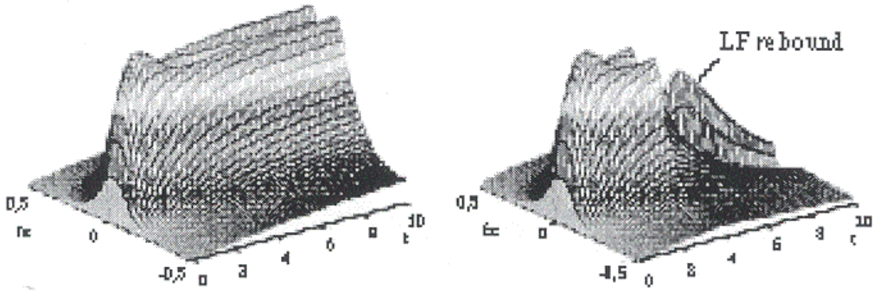

Fig. 1. Temporal evolution of the retina spatial transfer function, for a sustained stimulation (left) and for a stimulation of 40 ms duration (right). In this case, the low spatial frequencies which naturally decreased after 15 ms, reappear just after the end of the stimulation. This visual rebound could be attributed to the visual after-effect phenomenon.

From a theoretical point of view, this remanant information might link continuous stimulations together, and enhance coarse information processing, when we close our eyes, or after saccades, eye blinks or any sudden visual change in the world.

Consequently, the low level visual system represented by the retina naturally favours low spatial components analysis *nearby context changes*. This observation taken together with psychophysical studies strongly suggests that the low resolution structural information highlighted at an early level might be useful for further preattentive processing : to rapidly classify images, for example as outdoor or indoor scenes, or more precisely, as a specific scene category (as room, city, beach etc...). The similarity in global structure of images belonging to the same category might allow our preattentive visual system to automatically compute global perceptive correlations between components, independently of the exact local correlation between edges (see Fig.2).

In the next section, we present a new strategy for scene classification by an artificial system, which takes into account the global correlations in low and very low frequency ranges between real world photographs.

This kind of information may prove to be sufficient for a self-organising neural network [4] to easily cluster the images of the database into categories that make sense.

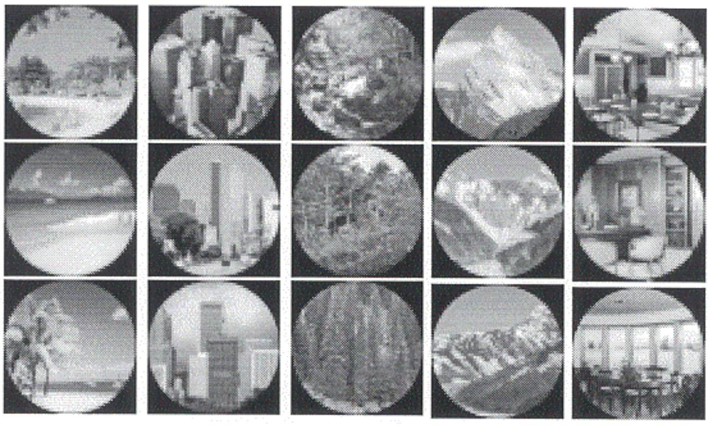

Fig. 2 Examples of photographs from different categories (beach, city, forest, mountain and room) used in experiments with human subjects and then automatically classified by a self-organising network (CCA).

4- Scene classification by CCA

A global structure analysis needs to be invariant with regard to the scene viewpoint and the precise location of the objects within the scene. Low frequency components of the power spectrum of the image are a good candidate to code this structural information. We propose to pave the spectral domain of an image by a set of oriented band-pass filters, with different centre frequencies, just like simple and complex cells present in striate cortex of vertebrates. These filters are of Gabor type, in phase or in quadrature, with centre frequencies f_0 , 2 f_0 etc... on an octave basis, and bandwidths such that the spectrum of the filters overlap at half height. The centre frequencies are 1/64, 1/32, 1/16 and 1/8 of the maximum sampling frequency (corresponding to twice of the spectral resolution of an image observer). The orientations of the filters are chosen in four directions: 0°, 45°, 90° and 135°.

In order to simulate complex cells, the squared outputs of in-phase and quadrature filters are added to obtain the power of the image in each of the 16 frequency domains (four frequency bands and four orientations). Therefore, an image is characterised by a 16-dimensional vector, whose components are represented by the contributions of its power spectrum to each of the 16 spatial frequency domains.

One should notice here two important points:

i- the centre frequencies of the filters range from *very low* (1/64) to *low frequencies* (1/8), disregarding the higher frequencies which are most relevant from image details (and local object recognition).

ii- the image database consists of 60 (256 grey-levels values) photographs. The scenes belong to 5 categories (beach, city, forest, mountain and room) and are selected from a large set of pictures, such that they represent *large within-class differences* and variability (see an example in Fig.2). These pictures represent scenes from different viewpoints, but common to all is that they have been easily recognised by human subjects as belonging to one of the five categories, when presented during a very short time (50 ms).

Every image of this data bank is represented by a point in a 16-dimensional space. We expect that the points corresponding to the same class of images (scene) will cluster in the same region of this space.

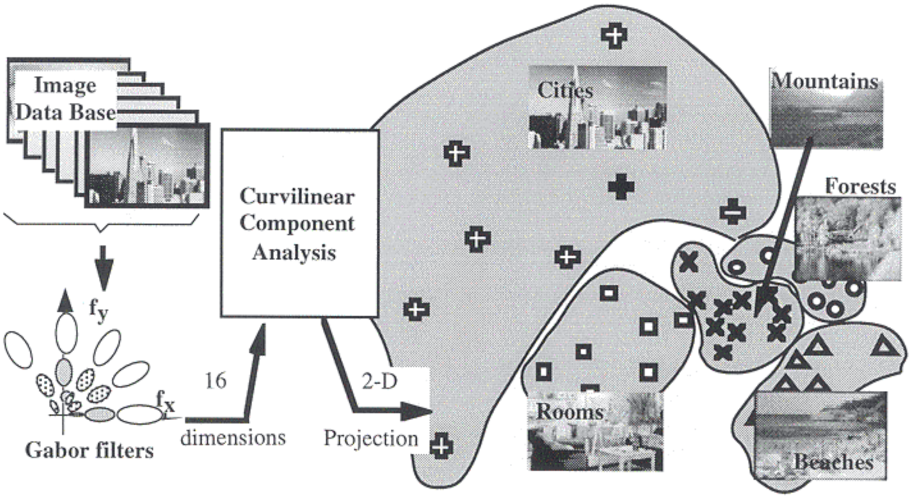

Fig. 3. The images are filtered by a bank of 16 oriented band-pass filters. Their clustering by the unsupervised CCA network provides classes, the neighbourhood of which being close to the classification naturally done by a naive human operator.

In order to compare the relative positions of the various clusters in the 16-dimensional space, we exploit the properties of a new self-organising neural network named CCA (Curvilinear Component Analysis) [4]. This network proceeds first, if necessary, to a vector quantisation of the 16-D input space, paving it with N prototypes in the submanifold spanned by the data. With a only 60-elements data base, this step is not necessary here. Then, it computes the non-linear mapping of the input space into a lower-dimensional output space Y, preserving the neighbourhood relationships between prototypes in the input space by copying them to the output space. The characteristics which make this algorithm outperform all other known algorithms that try to do a similar function are the following:

i- the cost function E to be minimised is the sum of differences of interpoint distances between prototypes, X_{ij} in input space and Y_{ij} in output space: $(X_{ij} - Y_{ij})^2$, weighted by a monotonically decreasing function *of the output distance* $F(Y_{ij})$. This contrasts with a classical non linear mapping algorithm. By applying this cost function the mapping is correct for small distances between prototypes in the output space, but not necessarily for small distances in input spaces: this allows the correct *unfolding* of the input submanifold for a suitable low-dimensional representation.

ii- In contradiction to Kohonen's Self-Organising Maps, this algorithm does not impose any *a priori* shape of the output data distribution. It organises itself according to the very shape and the interpoint distances existing inside the input submanifold.

In our experiment, the output dimension, revealing the one of the input data, reduces to 2 (see Fig. 3). In this output space, the interpoint distances copy those of the input space and reveal several clusters of points, each of them being associated with a specific category of scene. As a result, one can remark that:

i- the spreading of each class reveals the variability of images in the same category, the relative positions of clusters look like a similarity measures of classes.

ii- Artificial (cities, rooms) and natural (mountains, forests, beaches) scenes constitutes two super-clusters.

It is also interesting to notice that the 2-D representation which has been done by the network itself, without any supervised learning, clearly clusterises the samples of the data base into a particular *semantic* representation of the scenes. This strikingly corresponds to the type of responses human operators give when they are presented with the image as a flash of 50 ms: for example, under these conditions, naive operators are not able to recognise whether a particular scene represents Glasgow, Grenoble or Brugge, but will answer "city" (see [13]).

4- Conclusion

The presentation of this paper has been made slightly provocative, our aim was to strike the reader in order to make him think about a new idea.. In fact, the results, though interesting, are still preliminary. For example, the 2-D representation of a 16-D space is minimal, a 3-D one would provide better separations between classes, introducing colour information would greatly help clusterisation, some preprocessing of the data (reduction of zooming effects) would help in reducing the variability, then the spreading of clusters, more frequency bands scaled to half octaves would give better accuracy... Nevertheless, the principle of scene categorisation based on low and very low spatial frequencies seems to be shared by our artificial neural network and the behaviour of human, at a pre-attentional level and without any pre-assigned task.

Acknowledgements: The authors would like to thank Philippe Schyns, Pierre Demartines and William Beaudot for fruitful discussions and cooperations, and Stephane Perrier for his efficient programming of the Gabor filter banks.

5- References

[1] **Beaudot, W**. Le traitement neuronal dans la rétine des vertébrés: un creuset d'idées pour la vision artificielle. PhD from the Polytechnique National Institute of Grenoble, (1994).

[2] **Biederman, I., Mezzanotte, R.J., & Rabinowitz, J.C**. Scene perception: detecting and judging objects undergoing relational violations. Cog. Psychology, 14, 143-177 (1982).

[3] **DeValois, R.L., Albrecht, D.G., & Thorell, L.G.** Spatial frequency selectivity of cells in macaque visual cortex. Vision Research, 22, 545-559 (1982).

[4] **Demartines, P., & Herault, J.** Curvilinear Component Analysis: a self-organizing neural network for nonlinear mapping of data sets. IEEE Trans. on Neural Networks, 8 (1997).

[5] **Dong DW, Atick JJ.** Statistics of natural time-varying images, Network: Computation in Neural Systems, 6:159-178 (1995).

[6] **Henderson, J.M.** Object identification in context: the the visual processing of visual scene. Canadian Journal of Psychology, 46 (1992).

[7] **Herault, J. Jutten, C.** Reseaux Neuronaux et Traitement du Signal. Hermes, Paris (1994).

[8] **Jain, A.K., & Vailaya, A.** Image retrieval using color and shape. Pattern Rec., 29, 8, 1233-1244 (1996).

[9] **Mandler, J.M., Parker, R.E.** Memory for descriptive and spatial information in complex picture. Jl of Experimental Psychology: Human Learning and Memory, 3, 386-396 (1976).

[10] **Oliva, A., & Schyns, P.G**. Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. Submitted for publication.

[11] **Potter, M**. Meaning in visual search. Science, 187, 965-966 (1975).

[12] **Schyns, P.G., & Oliva, A**. From blobs to boundary edges: Evidence for time-and spatial-scale scene recognition, Psychological Science, 5, 195-200 (1994).

[13] **Schyns, P.G., & Oliva, A**. Flexible, rather than fixed, scale selection in scale and face recognition, Perception, in press.

[14] **Van der Schaff, A., & van Hateren, J.H**. Modelling the power spectra of natural images: statistics and information. Vision Research, 36, 2759-2770 (1996).