Multiscale feature extraction from the visual environment in an active vision system

Y.Machrouh[1], J.-S.Liénard[1], P.Tarroux[1,2]

Abstract. This paper presents a visual architecture able to identify salient regions in a visual scene and to use them to focus on interesting locations. It is inspired by the ability of natural vision systems to perform a differential processing of spatial frequencies both in time and space and to focus their attention on a very local part of the visual scene. The present paper analyzes how this differential processing of spatial frequencies is able to provide an artificial system with the information required to perform an exploration of its visual world based on a center-surround distinction of the external scene. It shows how the salient locations can be gathered on the basis of their similarities to form a high level representation of the visual scene.

Introduction

The use of active mechanisms seems to be a way to improve the abilities of machine vision systems. Active systems search salient features in the visual scene through a dynamic exploration. They can direct their search toward the most meaningful stimuli using attentional mechanisms leading to a reduction of the computational load [1,2].

Thus, natural vision is a behavioral task, not a passive filtering process. An exploration of the visual world that relates perception and action allows to label the external space with natural landmarks associated with the exploratory behavior. In this respect, the relationships between agents and natural systems suggest that certain aspects of natural perception can be successfully incorporated in artificial agents.

Otherwise, during the past few years, several studies have been devoted to the understanding of the essence of vision considered as an information processing mechanism [4]. This approach is grounded on Barlow's proposal [5] which stated that the main organizational principle in visual systems is the reduction of the redundancy of the incoming stimuli.

These considerations, issued form information theory, led several authors to analyze the statistical organization of natural images. They demonstrated that natural images (those which do not exhibit any specific bias in their pixel distribution) have a stationary statistics and an auto-similar structure. As a consequence of these characteristics, their power spectra fall off as $1/f^2$ [8].

In this context, different authors [6,14] demonstrated that a way to transform the initial redundancy was to improve the statistical independence of the image descrip-

[1] LIMSI-CNRS BP 133 F-91403 Orsay Cedex

[2] ENS 45 rue d'Ulm F-75230 Paris cedex 05

tors. According to this hypothesis, an image can be viewed as a linear superposition of several underlying independent sources.
The filters that provide this statistical independence can be computed through the application of the source separation adequate algorithms (Infomax, BSS, ICA).
One can show [6,14] that the optimal filters computed according to these principles are multiscale local orientation detectors similar to a Gabor wavelet basis [7].
However, although a lot of work has been devoted to the understanding of these theoretical bases of information processing in natural visual system, few attempts have been made thus far to use these principles in artificial vision systems. Practical implementations impose some limitations that require to analyze what is really obtained with simplified models based on these general principles. On the other hand, no artificial vision system has been designed to include both multiscale wavelet analysis and differential spatial and temporal processing of spatial frequencies. A prerequisite to the design of such a system is to be able to characterize the information obtained from a bank of wavelet filters in different frequency channels.
We thus analyzed here the information issued from various combinations of high and low frequencies of statistically uncorrelated signals. Our aim was to determine how to build a multivariate representation of the scene that allows a dynamic grouping of image points on the basis of their similarities in a given context and for a given task.

System overview

Image data

A set of 11 natural images selected from a larger database was used in the present study. Pictures that include too many traces of human activity (buildings, roads…) were avoided. Only images with similar initial resolution (around 256x512 pixels) were retained.

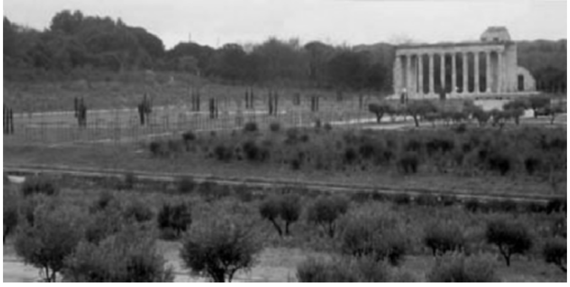

Figure 1. Sample image from the set of natural images used in the present work.(original size 512x256)

The images were discarded when their power spectrum did not fit the $1/f^2$ characteristics [8]. Figure 1 shows one typical example of an image used in the present study.

Initial filters

A guideline for this work was to retain among the filtering characteristics of the primate visual system those which can be useful for the elaboration of an artificial system of situated and active vision.
Two characteristics have attracted our attention: the elimination of image redundancy in the processing steps designed to maximize the statistical independence of the scene descriptors and the differences in the processing of spatial frequencies between the center and the surround of the visual field.
The visual scene was filtered by a first bank of Gabor wavelets in four spatial orientations and four spatial frequencies (1/8, 1/16, 1/32, 1/64). For each initial image we got 32 resulting images (two for each quadrature pair of each of the 16 Gabor filter). This multiscale processing was implemented using a Burt pyramid according to the method proposed by Guérin-Dugué [10].
For the purpose of this study and in order to obtain a complete view of what information is obtained from a detector during a systematic exploration of the visual scene, the whole scene was filtered by the entire bank of filters. In an operational system with a focal vision only a small part of these computations are needed.

Simple cells – Complex cells

An important distinction between the use of wavelets in image processing and the filtering steps in the visual system is the presence of strong non-linearities in the latter. Primary visual cortex shows several cell types according to the non linearities they implement. Simple cells (SC) perform an additive combination of their inputs. They respond to an oriented stimulus localized at the center of their receptive field. The so-called complex cells (CC), on the contrary, exhibit a kind of translational invariance and respond to a stimulus whatever its position in the receptive field of the cell.

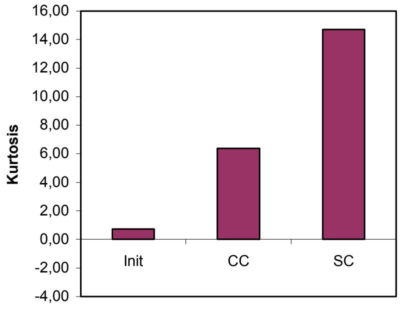

Figure 2. Effects of filtering of the statistical independance criterion. Init: Initial image, SC: Simple Cells, CC: Complex cells

Other cell types (mainly in extrastriate cortex) combine these outputs in order to be sensitive to curvature and terminations (end-stop cells).

To model simple cells we used additive units with a zero threshold ramp transfer function which amounts to take into account only the positive part of Gabor filters. The inhibitory part is indeed not transmitted by these cells.

According to Field [5], we modeled complex cells output as the norm of quadrature pair Gabor filters. We verified that this implementation effectively leads to a reduction of the redundancy for both cell types by a comparison of the kurtosis before and after filtering (Figure 2). Kurtosis is indeed a good measurement of the statistical independence of a set of detectors [9].

A third type of detector with large receptive fields and designed to provide a contextual information will be considered in the following section.

In order to build a set of higher level detectors suitable for the extraction of complex features we performed a Karhunen-Loeve transform of the outputs. A set of 1744 image patches (5x5) extracted randomly from the initial 11 natural images was used to build these spaces. We thus obtained 8 eigen-vectors at the output of simple cells and 4 eigen-vectors at the output of complex cells for each frequency band. These computations amount to a non-linear principal component projection of the initial image performed with two different types of non linearities.

Global energy – Local context

As stated above, we assumed the existence of detectors sensitive to the global energy in the different orientations. In an image region corresponding to the fovea, the system computes a global energy vector for each of the four orientations. This vector is used to build a signature that can be used to classify the region. Such an analysis provides us with contextual information [11,13]. We consider the identification of these contexts as a prerequisite for the recognition of objects. The importance of contextual information in natural systems can be deduced from the experimental observation that object recognition is effectively facilitated if the objects are viewed in congruent contexts [13].

Thus, the system computes three output sets on each image: (i) an output directly issued from the Gabor filters filtered by a ramp function (SC), (ii) an output giving the local energy at the output of these filters analogous to the output of complex cells (CC) and (iii) a large field output providing contextual information.

Results

Simple cells

For each image point the system provides a high dimensional vector made of 32 orientation components spread over 4 frequency bands for SC detectors and 16 orientation components in 4 frequency bands for CC detectors.

Although Gabor detectors maximize the statistical independence of their outputs, in practice they are not strictly independent. The analysis of these outputs through a

Karhunen-Loeve transform leads to a data representation basis that sorts the representations according to their greatest statistical significance.
The first axis corresponding to the highest eigen-value shows highly variable details from one scene to another (figure 3 left). It emphasizes details related to the structures present in the scene. This probably results from the fact that these structures are correlated in a given scene due to the correlation induced by the presence of objects. They are uncorrelated from one scene to another because each scene has a different organization.

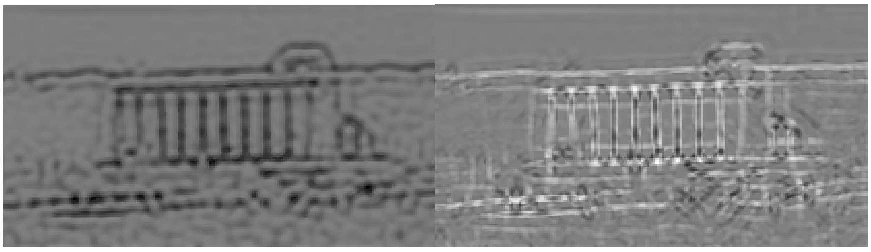

Figure 3. Output of SC filters: projection of the output along the first (top) and the last (botttom) eigen-vector of the output space

On the contrary, details filtered by the axes corresponding to the lowest eigen-values (figure 3 right) are expected to weakly contribute to the total variance. They correspond to features most frequently observed from one image to another.

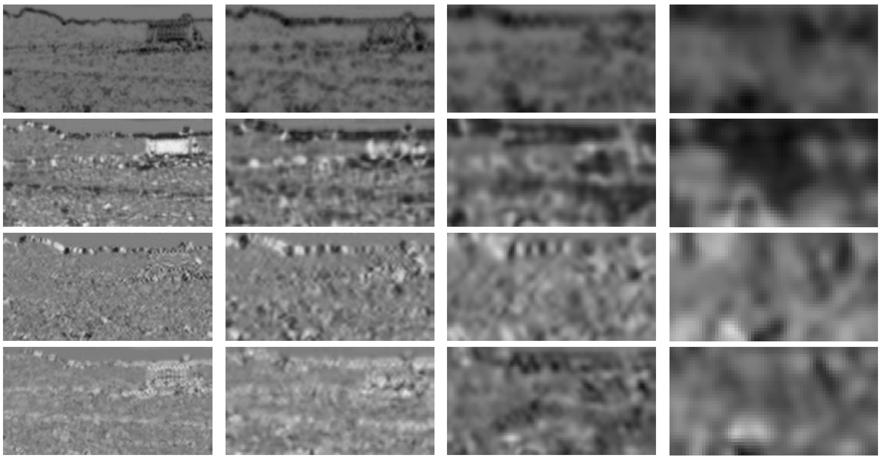

Figure 4. Eigen-images from CC filters. The images are computed as the projection of the CC outputs on the eigen-vectors defining the output space of these filters. Columns range from high to low frequencies (from left to right: 1/8 to 1/64). Lines show the filter outputs along the principal components (top: highest variance, bottom : lowest variance).

The same region revealed by the first projection axis (Figure 3 left)(% initial variance : 29.4%) of the KL transform and the last projection axis (Figure 3 right)(% initial variance : 2.47%) shows that, while the first axis tends to reveal long edges that contribute significantly to the general structure of the objects, the last axis tends to reveal termination and curvature points that are not characteristic of the image structure.

We obtain a complex set of features along the different axes. The most representative of the presence of objects correspond to the first axes. On the others, features representing complex combinations of stimuli frequently observed in natural images seem to be sorted according to their level of abstractness.

Complex cells

The same transform can be applied to the output of complex cells. Figure 4 shows the main axes of the KL transform following the computation of the Gabor norm for different spatial frequency bands.

The projection axes (rows in the figure) extract distinct features from the initial image as well within the same frequency band (rows) as between different frequency bands (columns)(note that for instance the building vanishes in axis 3 projection. Figure 4 3rd row). These features are entirely different from those extracted by the output transform of SC.

One can observe that high frequency details disappear in low frequency channels except for objects which exhibit frequency similarities (high frequency details repeated over a large area like the building).

Objects in the foreground, which are apparently characterized by low frequencies, appear in low frequency channels while they are not represented in high frequency band. Low frequency channels are able to distinguish features that have some spatial extension (the building or the foreground bushes).

A comparison of the lowest frequency channels (Figure 4 right column) shows that the locations revealed on the different axes are largely uncorrelated, thus corresponding to different points of view on the scene.

The lesser number of low frequency features (figure 4 right column) defines a small set of landmarks able to characterize the visual space and to guide exploratory saccades. This low-frequency information is the only one available in the periphery of the visual field.

Correlation between channels

One of the important questions raised by this analysis is how different are the indices obtained from the different frequency channels. If two channels correspond to the same combination of basic features, the corresponding eigen-vectors should be similar. Thus, a measure of the similarity between the eigen-vectors in different frequency bands is given by the product of the eigen-matrices in these frequency bands. Using this method we compared the output spaces of respectively simple and complex cells for different frequency bands. We obtained strongly different results for the comparison of output spaces in SC channels and in CC channels.

For simple cells, the correlation between the axes of the spaces corresponding to different frequencies are low and distributed over the different axes (data not shown) while in complex cells the respectively high and low frequency bands exhibit similarities (table 1).

Table 1. Analysis of the output space for CC detectors. The eigen-vectors corresponding to the same axes show a very high correlation between respectively high and low frequency channels. The cross-correlation between eigen vectors corresponding to different axes is usually low (not reprinted here)

	Frequencies					
Axes	f0/f1	f0/f2	f0/f3	f1/f2	f1/f3	f2/f3
F1	0.990	0.442	0.410	0,365	0,330	0,996
F2	0.997	0.517	0.501	0.507	0.486	0.997
F3	0.991	0.363	0.370	0.425	0.424	0.995
F4	0.994	0.656	0,653	0.641	0.630	0.996

These results lead to the conclusion that the combination of simple cells outputs across the frequency bands underline uncorrelated details, whereas the outputs in high (resp. low) frequency bands correspond most frequently to similar stimuli.

A pyramidal decomposition of the scene allows to combine these characteristics to identify spatial positions characterized by spectral compositions as diverse as possible.

This diversity seems to lead to a greater separability of these spatial positions and seems to be able to facilitate objet discrimination.

Identification of global contexts

Cells sensitive to low frequencies have large receptive fields. However in higher layers of the visual system cell types that encode intermediate representations also exhibit larger receptive fields. They combine the output of the cells in the preceding layers and gather the information coming from brighter regions of the visual field.

A vector that combines the global energy components associated with each frequency channel provides a suitable code for representing the whole fovea. It has been shown that such vectors can be used to classify visual scenes according to the context they belong to [11,13]. In the present study, we build such detectors in computing the mean energy provided by the output of CC cells in the four frequency bands already mentioned.

To determine how spatial indices provided by the channels previously described can be used for the identification of interesting locations in the scene, we performed the following experiment:

A set of salient locations are computed from the eigen-images defined previously. Points in the image are selected at random or on the basis of these salient locations. At each point the mean energies of the CC outputs in an image window corresponding to the fovea were computed for each frequency. We thus obtained an energy vector for each of the selected point. A PCA analysis was performed on this set of vectors. One should keep in mind that this use of PCA differs from its use in the previous sections. The Karhunen-Loeve transform was previously used as a self-organization tool leading to a set of linear combination defining complex features frequently occurring in natural images. In this section, PCA should be considered as a mean to analyze the structure of the space at the output of the SC and CC filters.

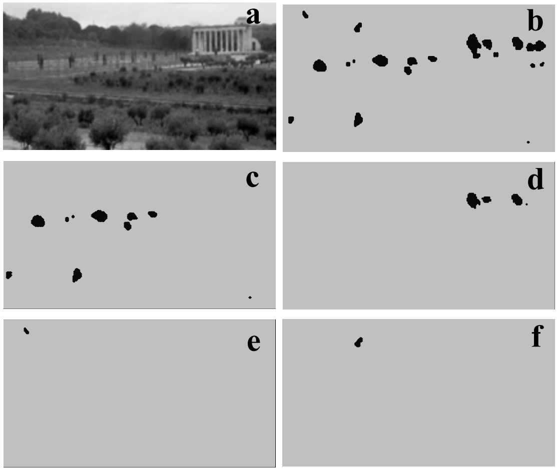

Figure 5. Clustering of fixation points corresponding to different regions of the visual scene. Clusters were identified on the first three principal components and the fixation points corresponding to each cluster plotted on the diagrams at their position in the initial image (a). (b) fixation points obtained from the second eigen-image and the second frequency channel shown Fig. 4. The other diagrams show the location of some clusters gathering salient points on the basis of their spatial frequencies and orientation properties: (c) trees and bushes, (d) building, (e) strong curvature at the border between hill and sky (f) another region of interest at the same limit

When the locations in the image are selected at random no obvious structure were observed in the PCA space. On the contrary, when they are selected on the basis of their saliencies, clusters were identified in the PCA space. Figure 5 show the locations of some of these clusters on the original image. Points corresponding to a similar context are grouped into the same cluster. The example shows for instance the ability of the method to separate fixation points on the basis of their natural or artificial nature (Figure 5 c and d).

It should be noted that Figure 5 shows only a small sample of the structures that can be identified. Only 1/16 of the available dimensions is presented here. Thus, the method transforms the initial image into a huge set of clusters each characterized by similar spectral signatures.

Discussion and conclusion

The visual filter system proposed in the present work produces a set of features that can be used to guide the exploration of the external scene. The features extracted by the non linear combination of SC channels seem rather suitable for object recognition. Features obtained from the computation of local energy (CC channels) allow a partition of the image into salient regions arranged according to their frequency composition. The computation of the global energy provides local context information and can be used to segment the scene on the basis of its spectral characteristics.

Thus, the output of this filtering system provides on one hand locations of interest able to guide an attentional system and on the other hand clusters of locations arranged according to their spectral signature.

This approach can be considered as an extension of textures segmentation methods [3] to the question of the identification of contexts and an extension of the method proposed by Hérault [11] to the analysis of local contexts. However it emphasizes the relativity of the context notion; the segmentation of the visual scene in (i) a global context and (ii) objects is an oversimplification

The visual scene is thus scattered into a set of projections on several disjoint subspaces. In each of these subspaces, salient points form clusters according to their similarities. These salient points are projected into disjoint sets of clusters and the corresponding objects can thus be grouped according to different points of view.

An object class is not characterized by a unique high level representation, but by the transient association of a subset of properties. This association can thus dynamically depend on the current task. Objects are not considered as similar and grouped on the basis of their intrinsic properties but according to those of their properties linked to a given goal.

A further step in this work will be to demonstrate how such coding abilities could indeed facilitate object classification. This requires to incorporate the present algorithms in the control architecture of a perceptive agent such that it can build a hierarchy of perception-action links based on the dynamic grouping of the perceived features.

ACKNOWLEDGMENTS

This work was supported by a grant from the "GIS Sciences de la cognition" CNRS.

REFERENCES

[1] Allport, A., Visual attention. In M.I. Posner (Ed.), *Foundations of cognitive science*, The MIT Press, 1989.

[2] Aloimonos, Y. (Ed.), *Active Perception*, Lawrence Erlbaum, Hillsdale,NJ, 1993.

[3] Andrey, P. and Tarroux, P., Unsupervised segmentation of Markov Random Field modeled textured images using selectionist relaxation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (1996) 252-263.

[4] Atick, J.J. and Redlich, A.N., Towards a Theory of Early Visual Processing, *Neural Computation*, 2 (1990) 308-320.

[5] Barlow, H.B., Possible principles underlying the transformation of sensory messages. In W. Rosenblith (Ed.), *Sensory Communication*, The MIT Press, cambridge, MA, 1961, pp. 217-234.

[6] Bell, A.J. and Sejnowski, T.J., The "independent components" of natural scenes are edge filters, *Vision Research*, 37 (1997) 3327-3338.

[7] Daugman, J. and Downing, C., Gabor wavelets for statistical pattern recognition. In M.A. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks*, The MIT Press, Cambridge, MA, 1995, pp. 414-420.

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

[9] Field, D.J., What is the goal of sensory coding?, *Neural Computation*, 6 (1994) 559-601.

[10] Guérin-Dugué, A. and Palagi, P.M., Implantations de filtres de Gabor par pyramide d'images passe-bas, *Traitement du signal*, 13 (1996) 1-11.

[11] Hérault, J., Oliva, A. and Guérin-Dugué, A., Scene categorisation by curvilinear component analysis of low frequency spectra. , *ESANN'97*, Bruges, 1997, pp. 91-96.

[12] Linsker, R., Self-organization in a perceptual network, *Computer Magazine*, 21 (1988) 105-117.

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

[14] Olshausen, B.A. and Field, D.J., Emergence of simple-cell receptive field properties by learning a sparse code for natural images, *Nature*, 381 (1996) 607-609.