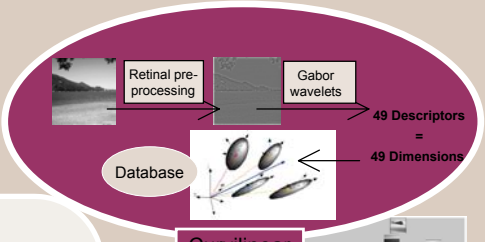
Natural Scene Categorization: Human vs. Machine.

Guyader Nathalie[1,2], Alan Chauvin[1,2], Le Borgne Hervé[1] {nguyader,hleborgn,alan.chauvin}@lis.inpg.fr,
[1]LIS, 46 Av. Félix-Viallet, 38031 Grenoble Cedex, France
[2]LPE, BP 47, 38040 Grenoble Cedex 9, France

Introduction

The aim is to optimize a machine-based semantic categorization of natural images according to human perception categorization.



Biological Model

Retinal pre-processing :
The retinal photoreceptors make a spatial high-pass filtering after an adaptive compression process. This results 1/ in a contrast equalization over the whole image and 2/ spectral whitening which compensates for the 1/f image amplitude spectrum [1].

Cortical filtering :
In area V1 of the visual cortex, the retinal image is decomposed by the filtering of cortical neurons, which are sensitive to various spatial frequency bands and various orientations of the stimuli. Here, we aim at categorizing and not describing scenes, so we simulate complex cells which are invariant to object position in the scene. These cells, described by means of Gabor wavelets, provide a local energy spectrum of images.
Images are filtered by Gabor wavelets into 7 frequency bands and 7 different orientations. According to the physiological data about the visual cells [2], the relative radial bandwidth of the Gabor filters is fixed at 1 octave. Each image is analyzed with a bank of 49 Gabor filters : it is then represented by a point in a 49-dimensions space whose components are the output energies of filters. Figure 1 is a special non-linear MDS representation (CCA [3]) of the image database obtained from interpoints distances.

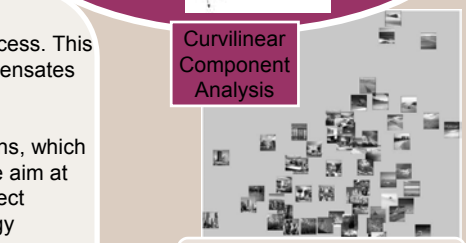


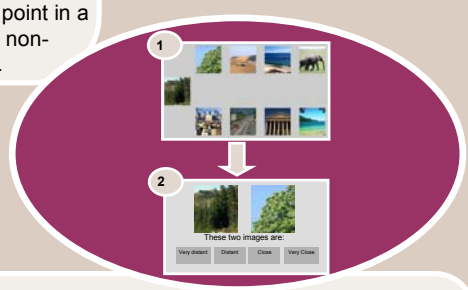
Figure1: *CCA-2D representation of the database for the Biological Model.*

Human Categorization

In this experiment, human observers judge the similarities between 105 selected images.

Participants : 48 subjects (ages between 18 and 50, with normal or corrected vision) were volunteers to participate in the experiment.

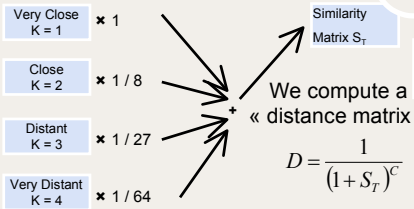
Stimuli : the image database covers a wide range of natural environments (same types as in [4]): animals, people, indoor scenes, nature as beaches or mountains…. Images are luminance ones; the size of each image was approximately 5,3 × 5,3 centimeters and subtended approximately 5 degrees of visual angle.



Protocol :
1st : the subject has to choose which image, among the eight images on the right side, is the most similar to the referenced one. (time limited to 5s.)
2nd : the subject has to explicit the proximity of the selected image to the reference one, on a scale of four level which are: “very close”, “close”, “distant”, “very distant”.

Computing the similarity matrix:

For four level of similarity *K* (K=1,…,4), we compute an elementary matrix S_K, increasing $S_K(i,j)$ of one unit each time a subject associates a test image *j* to a reference image *i* and chooses the level of similarity *K*. Thus the overall similarity matrix is a weighted accumulation of this four matrices. It is further transformed into a distance matrix



Curvilinear Component Analysis

Figure 2: *CCA-2D image organization for the human perception matrix.*

Model Optimisation

We want to take into account the human perception in our model, for that we minimize the following cost function of the weight ω_k:

$$C = \sum_{i,j} \left(Deuc_{ij}^2 - E_{ij}^2\right)^2 = \sum_{i,j} \left(Deuc_{ij}^2 - \sum_{k=1}^{49} \omega_k \left(data(i,k) - data(j,k)\right)^2 \right)^2$$

where *Deuc* is the Euclidean distance matrix of the 2D human perception space and *E* the Gabor model ones. *Data(i,k)* is the k^{th} filter component of the i^{th} image. The minimization of this function provides a weighting vector $\Omega = (\omega_1, \ldots, \omega_{49})$.
The improvement of our model is measured by the increase of the correct recognition rate.

For that we use a simple classifier: the mean vector of each category is computed, then we measure the Euclidean distance between each image and the different mean vectors. Then each image is associated with the “nearest” category. With this image database, we increase the percentage of correct categorization by 10% (See for example Figure 3).

Figure 3: *On the left, a zoom of the representation with Gabor filter descriptors, and on the right, a zoom of the same region projected by weighted Gabor filters.*

[1] J. Hérault, “De la rétine biologique aux circuit neuromorphiques”, *in Trait. IC2, Les Systèmes de Vision, J.M Jolion ed. Hermès*, 2001.
[2] R. L. De Valois & K. K. De Valois, “Spatial Vision”. *Oxford Univ. Press*,1988.
[3] P. Demartines & J. Hérault, “Curvilinear Component Analysis: a Self-Organising Neural Network for Non-Linear Mapping of Data Sets “, *IEEE Trans. on Neural Networks*, 8, 1, 148-154,1997.
[4] A. Mojsilovic and B. Rogowitz, “Capturing image semantics with low-level descriptors”, Proc of ICIP’01, vol 1, pp 18-21, Thessaloniki, Greece, 2001.