Towards the introduction of human perception in a natural scene classification system

Guyader Nathalie[1], Le Borgne Hervé[1], Hérault Jeanny[1] and Guérin-Dugué Anne[2]

[1]LIS, 46 Ave. Félix-Viallet, F-38031 Grenoble Cedex, France
[2]CLIPS, Bat B, rue de la bibliothèque, 38041 Grenoble Cedex 9, France
{nguyader, hleborgn}@lis.inpg.fr, Jeanny.Herault@inpg.fr, Anne.Guerin@imag.fr

Abstract : In this paper we develop a method to optimize a machine-based semantic categorization of natural images according to human perception. First, the categories are determined through a psychophysical experiment. The similarity matrices obtained from human responses are analyzed by a multidimensional scaling technique called Curvilinear Component Analysis (CCA). The same is done with an automatic image indexing system based on similarities between the outputs of Gabor filters applied to the images. Then we show that, by using the human categorization to balance the filter outputs, the system's performance may be significantly improved.

Keywords : Perceptual similarity, clustering, non-linear mapping, Gabor filter.

1. INTRODUCTION

The rapid growth of multimedia databases, and particularly digital image libraries, have created new needs for various users like publishers or journalists, criminologists (criminal identification), business people (trademark description), artists and teachers (encyclopaedias), that is persons who classify, organize, or navigate through databases. From the works made about textual databases, one could imagine to associate words with images and take advantage of existing text mining systems. However, two drawbacks limit such an approach. First, it would require a tedious manual indexing, while users would prefer an automation of this task. Second, it is known that a list of words would never exhaustively describe an image. That is why most researches on image databases have focused on the use of "low-level features" obtained from the raw pixel values [1]. For a complete review see [2, 3]. These techniques aim at indexing databases without any human intervention: some discrimination between broad classes of images have been made with success [4, 5, 6, 7], but difficulties are encountered to access at finer level of description [8]. Faced to a Content Based Image Retrieval (CBIR) system, a user tries to perform different kinds of tasks [1]. A classical approach would be the exploration of a database with a precise idea of the target image. However, one could imagine a journalist who does not have an exact idea of the image he searches, but only an idea of the image topic. At last, we may consider the case of a good knowledge of the searched picture, but the image does not exist in the considered database. In each case, we should focus on the importance of the visual *context* of the image, that is to say its category [9]. Categorization is a crucial first step for retrieving images, but low-level descriptors commonly used in image database navigation and retrieval do not catch the high-level semantic of images. More, few CBIR systems have taken into account the characteristics of human visual perception and the underlying similarities between images it implies.

In this paper, we develop a method which directly introduces human perception in an artificial model of vision. Our purpose is to provide a more user-intuitive image categorization. First, we describe a psychophysical experiment which is an improvement of the Rogowitz and al.'s "Computer Scaling" [10]. The results reveal a human perceptual space of natural scenes represented by a multidimensional technique called Curvilinear Component Analysis [11]. This organization allows us to identify the major semantic categories according to a human perceptual space. Finally, we use the knowledge directly extracted from this human categorization to drive a model of representation of image databases, based on Gabor filters. See [17] for another approach.

2. EXPERIMENT OF COMPUTER SCALING

In the experiment, human observers judge the similarity of 105 selected images presented on a computer display. We measure the perceived similarity of each image with every other image of the base. In each trial (Figure 1), a reference image is presented with eight randomly-chosen images; the subject is asked to choose which image, among this context, is the most similar to the reference one and to give the similarity level. This judgement of similarity in a second step did not exist in the original "Computer Scaling" experiment. It allows the subject to make "weak association" of images and "strong" ones. This experiment provides similarity matrices (one per similarity level). The parameters of 105 and 8 images will be justified later. In order to rely only on structural information, we did not consider color images.

2.1 Selecting the stimuli and subjects

The image database consists of a set of 105 natural images. These images have been selected in order to cover a wide range of natural environments (same types of scenes as in [12]): animals, people, indoor scenes, nature as beaches or mountains, buildings…. The experiment was conducted on a monitor (luminance TIFF images), using a Matlab interface. The display measured 36,5 × 27,5 centimeters and it is viewed at a distance of approximately 60 centimeters. Viewed on the display monitor, the size of each image was approximately 5,3 × 5,3 centimeters and subtended approximately 5 degrees of visual angle. Figure 1 shows one trial of the experiment.

We selected subjects who ignored the real purpose of the experiment. The results were computed with 48 subjects with normal viewing or corrected to normal viewing. At the end of each experiment, we collected some information from the subjects to identify their similarity criterions and to guide our experiment result analysis.

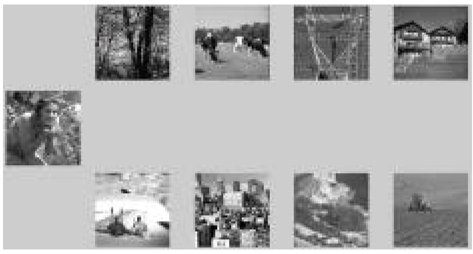

Figure 1 : One trial of the experiment. The subject has to choose which image, among the eight images on the right side, is the most similar to the referenced one (on the left).

2.2 Experiment description

In this type of experimental paradigm, measurements are made using paired image comparisons. In order to represent the range of possible natural scenes, we need a sufficient number of categories, as well as a sufficient number of images per category, while keeping the database size within reasonable limits for psychophysical experiments. So, we decided to work with approximately one hundred images. The comparison of every pair would take a too long time. So, one image is presented with eight-randomly chosen images among the 104 others, as in the Rogowitz's experiment [10]. To compare one image with the 104 others, 13 trials are necessary (13*8 = 104); so, to fill a complete similarity matrix, 1365 trials are needed. For one subject, the time required for all these trials is too long. Thus, the trial number is divided by 4; and the experiment for each subject lasts only thirty minutes. A "full" similarity matrix is built from four different subjects.

The experiment is organized as follows: subjects are placed in front of a display. In a first step, which corresponds to a first screen, the subject is asked to select, among eight images, the one which is the closest in term of similarity to the reference image. This first part of each trial is limited to five seconds, which is enough to glance at the eight presented images. Thus, the association is based on global criteria, which is consistent with our global model of Gabor filter description (see part 4). There is a second step consisting on a proximity judgement. The subject has to tell the proximity of the selected image to the reference one, on a scale of four levels. These levels are: "very close", "close", "different", "very different". Thanks to this second step, subjects can relativize their responses, if they judge that the closest image among the

eight proposed, is not so similar. They have as much time as they want to judge similarity. This step allows us to tell whether the best match was selected because it was really similar to the reference or because it was the less dissimilar.
Before the experiment, the subject is specified to look at all the presented images before choosing the more similar one and that their reaction time is not measured. We also tell that we do not care for the “non clicked” trials. Here, we have 12 “full” matrices (that is 48 subjects). We consider the ordering given by the human subjects as a measurement of perceived similarity. We expect that our experiment show a scene organization into clusters, one cluster being ideally representative of one semantic category.

3. DATA ANALYSIS

The first step in the data analysis is to compute the similarity matrix from the responses of human subjects. In order to use it as the entry of our multidimensional scaling data analysis algorithm, we transform this matrix into a distance one. This algorithm allows us to project images into a 2-dimensions space which enhances the meaningful image categories, at least those which would be recognized in a Content Based Image Retrieval (CBIR) paradigm.

3.1 Computing the Similarity Matrix

The overall similarity matrix $\mathbf{S_T}$ is the average of the sample similarity matrices $\mathbf{S}$ obtained by a weighted accumulation of the elementary matrices $\mathbf{S}_K = \{S_K(i,j)\}$, one for each of the four levels of judgements (matrix $\mathbf{S_1}$ for “very close”, $\mathbf{S_2}$ for “close”, $\mathbf{S_3}$ for “distant”, $\mathbf{S_4}$ for “very distant”). Each time, a subject associates a test image *j* to a reference image *i*, we increase $S_K(i,j)$ of one unit. In fact, the choice of the relative weights for the accumulation has been done considering that it exists a non-linear relation between perceived proximity, and judged similarity. Generally, if A and B are the representations of the stimuli A and B in the feature space, then *d(A,B)* is the perceptual proximity, while the judged similarity is

$$d\,(A,B) = g(d(A,B)) \qquad (1)$$

where g, is a suitable monotonically non-decreasing function of its argument [13]. In our case, we assumed a relation between $\boldsymbol{d}$and the perceived distance *d* as:

$$\boldsymbol{d} = d^{\frac{1}{3}} \qquad (2)$$

This relation expresses the good visual discerning for short perceptual distances, and the human tendency to mix the large and very large perceptual distances in their judgement. Assuming that, the similarity between two images is the inverse of the perceived distance, we have chosen the following weighting for **S**:

$$S(i,j) = \frac{S_1(i,j) + \frac{1}{8} S_2(i,j) + \frac{1}{27} S_3(i,j) + \frac{1}{64} S_4(i,j)}{(1 + 8 + 27 + 64)} \qquad (3)$$

The denominator is a factor which normalizes the entries of **S** so that they vary from 0 to 1. Now, the overall similarity matrix $\mathbf{S}_\mathrm{T}$ is the average of the sample similarity matrices **S**.

3.2 Matrix of distance

Most of the multidimensional scaling algorithms such as Curvilinear Component Analysis (CCA) use dissimilarity matrices, so we have to transform $\mathbf{S}_\mathrm{T}$ into a “distance matrix” **D**. The sparsity of **S** is 50% for the 12 matrices of responses. It is worth stating that most of the entries of $\mathbf{S}_\mathrm{T}$ have small or null values. That is the reason why we decide to transform its entries through a non-linear function *D(.) = 1/S(.)*, to set out small values, rather than to use a classical difference *D(.) = 1-S(.)*. Furthermore, we want a normalized distance between [0, 1]; that gives the following equation:

$$D(i,j) = \frac{\frac{1}{\left(1 + S_T(i,j)\right)^C} - 2^C}{1 - 2^{-C}} \qquad (4)$$

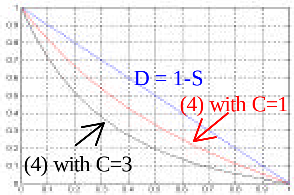

Figure 2: From S to D.

It brings a larger variance of distance for the small values of $\mathbf{S}_\mathbf{T}$ (figure 2). The variance of distances are adjusted with the coefficient *C*. In fact, **D** is not symmetric since the value of *D(i,j)* designed the distance between images *i* and *j*, when image *i* was the reference image. We measure the “percentage” of non symmetry with the variable:

$$X_{ij} = \frac{\left|D(i,j) - D(j,i)\right|}{D(i,j) + D(j,i)} \qquad (5)$$

This variable is belongs to [0, 1]; its mean, over **D**, is 0.03 which is very low. Considering the properties of distance matrix, we compute our final distance matrix as the mean between itself and its transpose; the result is then a symmetric matrix. This operation rules out the semantic interpretation of non-symmetries, but this fact will be considered in future works.

This symmetric matrix is then processed by a CCA which allows us to project the human judgement into a 2-dimensions metric space (see Figure 4, part 5). It is clear that these results suggest semantic categories. The information about how these clusters are formed will be used in order to tune our computer-based system for image categorization.

4. GABOR FILTERING MODEL

Former studies have shown that the global distribution of the local dominant orientations appears to be a powerful feature for discriminating between four semantic categories of real world scenes (urban scenes, indoor scenes, open landscapes and closed ones) [9]. Our model is inspired by the biology of the primary visual cortex (bench of oriented band-pass 2D filters). In order to cope with the global characteristics of an image, we work with the total energy distribution according to spatial frequencies and orientations.

This section explains how we compute the different processes of human visual system. After a retinal pre-processing (adaptive equalization of local contrasts and spectral whitening), a "cortical" filter, which codes images by means of its energy within seven spatial frequency bands and seven orientations, is applied. Each image is then coded by a 49-dimensions vector.

4.1 Retinal pre-processing

Our visual system architecture is an interesting model because it is known to possess functions able to get rid of some variabilities which would impair image classification. The retinal photoreceptors make a space-time high-pass filtering after an adaptive compression process (Figure 3). It results in a contrast equalization of the image (from where a relative insensibility to illumination variations) and a spectral whitening which compensates for the 1/f image amplitude spectrum [14].

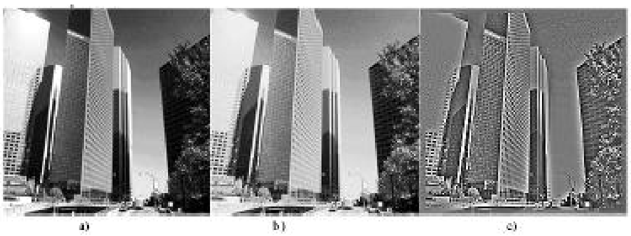

Figure 3 : Retinal pre-processing. The original image (a) is subject to a local adaptive compression by photoreceptors (b), then is high-pass filtered by the retinal circuits (c). Notice contrast equalization across the whole image.

4.2 Cortical filtering

In the V1 area of the visual cortex, the retinal image is decomposed into a certain number of primitives by the filtering of cortical neurons, which are sensitive to various spatial frequency bands and various orientations of the stimuli. There are two types of cortical cells: simple cells which can be simulated by in-phase and in-quadrature Gabor filters and the complex cells which integrate the visual stimuli energy [15]. 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 the means of Gabor wavelets, provide the local energy of images.

Images are filtered by 49 Gabor wavelets into 7 frequency bands and 7 different orientations. According to the biological data about the visual cells [16], the relative radial bandwidth of the Gabor filters is fixed at 1 octave. As we want to have the whole image characteristics, we compute a bank of Gabor filters which covers the spectral domain as well as possible. Having chosen 7 orientations, we take a transversal bandwidth of 180°/7. The frequency spectrum is log-polar sampled: the center frequency, f_k , of the k^{th} filter is : $f_k=1.5*f_0$, where f_0 is the center frequency of the lowest frequency filter. So, each image is analyzed with a bank of 49 Gabor filters, the output energies of which provide a point in a 49-dimensions space.

5. Results

It is known that some semantic categories of real world scenes as indoor and outdoor scenes are revealed by the study of the energy distribution [9]. The computer scaling experiment emphasizes these categories, it shows new categories as people or animals and splits some into sub-categories; for example it provides a separation in the beach cluster between scenes with only a beach and scenes with a beach and buildings or others “objects”. With our classification model, we obtain a distribution of scenes, not as well clustered as by the human perception. For image indexing or image database navigation, it would be better to have a more clustered image organization; that is why we decided to drive the scene organization obtained with our Gabor filtering model by some information derived from the human perceptual organization.

5.1 Finding a human perception space

We take the distance matrix **D** obtained by the measures described in part 3. This matrix is neither related to any known space, nor to any known metric. A number of methods provide a representation of this unknown space. Here, with the CCA, the local topology of the input average manifold contained in the distance matrix is mapped into a 2-dimensions representation space (Figure 4). We can distinguish semantic clusters in the data, like indoor scenes, people, forests, deserts, mountains, buildings, animals, fields.

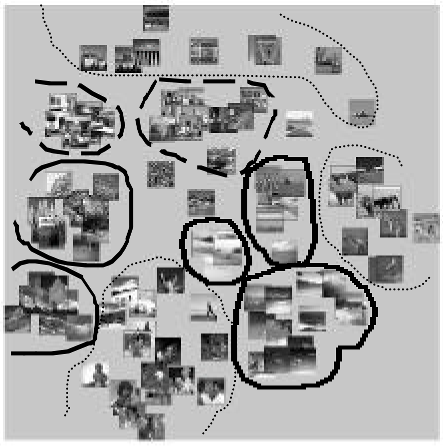

Figure 4: 2-dimensions image organization with human perception matrix.

We take into account this 2D representation of scenes to compute a dense Euclidean distance matrix, **Deuc**, between images. Then, we fit this new matrix with the Euclidean distance matrix between our global image descriptors (Gabor filters).

5.2 Results and notes

We minimize the following cost function of the weight ω_k:

$$C = \sum_{i,j} \left(Deuc_{ij}^{2} - E_{ij}^{2} \right)^{2} = \sum_{i,j} \left(Deuc_{ij}^{2} - \sum_{k=1}^{49} w_{k} \left(data(i,k) - data(j,k) \right)^{2} \right)^{2} \quad (6)$$

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 (6) provides a weighting vector $\Omega = (\omega_1,\ldots,\omega_{49})$.

First of all, it is important to note that our model cannot discriminate classes such as people or animals because the used descriptors are global ones; So, a little change in the context as "objects" cannot be detected by the study of the global spectrum. In the same way, people or animals do not have particular direction in their signal, so they cannot be discriminated using Gabor descriptors. The study of these particular scenes will be for future work.

The Gabor representation of images is less clustered than the human representation, but, a clear separation appears between natural and artificial scenes or between open and close landscape. Meanwhile, some images as trees with vertical directions are classified as cities, because of our descriptors, which detect merely spatial frequencies and orientations. Due to this fact, we apply the weighting vector Ω on our filters before projecting by CCA into a 2D space. This weighting reveals to largely improve the categorization and the semantic clusters appear more marked.

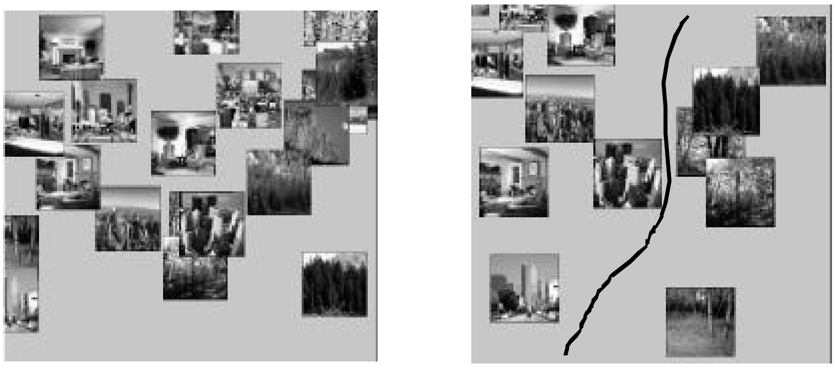

Figure 5: On the left, a zoom of the representation with Gabor filter descriptors, and on the right, a zoom of the same region projected by weighted Gabor filters.

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%.

6. CONCLUSION

In this paper we have introduced a mean to directly take into account the human perception in a classification image model. Using a non-linear mapping on similarity matrix provided by an image similarity experiment, we have found a human perception organization subspace for natural scenes. It is well known that a Gabor-based vision model could well enhance the structural information detected by complex cells of the visual cortex. By using data from the psychophysics experiment, we have modified the relative importance of the features. This brings significant results since the model exhibits more "intuitive" organization, by clearly grouping semantic categories.

7. ACKNOWLEDGEMENT

The authors would like to thank C. Berrut from CLIPS, and A. Chauvin, S. Donadieu, C. Marendaz and C. Peyrin from LPE for fruitful discussions. This work is partially funded by ELESA and IMAG. The Rhône-Alpes region funds H. Le Borgne.

8. REFERENCES

[1] I.J. Cox, M.L. Miller, Omohundro, P.N. Yianilos. “PicHunter: Bayesian Relevance Feedback for Image Retrieval”, Int. Conf. On Pattern Recognition, Austria, 1996.

[2] B. Johansson , “A Survey on: Contents Based Search in Image Databases”, 2000. http://www.isy.liu.se/cvl/Projects/VISIT-bjojo/survey/surveyonCBIR/index.html,

[3] A. Del Bimbo, “Visual Information Retrieval”, M. Kaufmann Ed, San Francisco, USA, 1999.

[4] A. Guérin-Dugué, A. Oliva, Classification of Scene Photographs from Local Orientations Features, Pattern Recognition Letters, 21, pp 1135-1140, 2000.

[5] Gorkani M.M., Picard R.W., Texture Orientation for Sorting Photos at Glance, IEEE conference on Pattern Recognition, octobre 1994.

[6] Szummer M., Picard R.W., Indoor-Outdoor image classification, IEEE Int. Workshop on Content-Based Access of Image and Video Database / ICCV’98, 1998.

[7] Vailaya A., Jain A., Zhang H.J., On image classification : City images vs Landscapes, Pattern Recognition, vol. 31, n°12, pp. 1921-1935, 1998.

[8] I.J. Cox, Ghosn J., M.L. Miller, T.V. Papathomas T.V., P.N. Yianilos, “Hidden Annotation in Content Based Image Retrieval”, Proc of CVPR’97, 1997.

[9] J. Hérault, A. Oliva and A. Guérin-Dugué, “Scene Categorisation by Curvilinear Component Analysis of Low Frequecy Spectra”, ESANN’97, Brugge, April 1997.

[10] B. Rogowitz, T. Frese, J. Smith, C.A. Bouman, and E. Kalin, “Perceptual image similarity experiments”, Human Vision and Electronic Imaging III, *Proc. of SPIE*, vol. 3299, San Jose, CA, January 26-29, 1998.

[11] P. Demartines and 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.

[12] A. Mojsilovic and B. Rogowitz, “Capturing image semantics with low-level descriptors”, Proc of ICIP’01, vol 1, pp 18-21, Thessaloniki, Greece, 2001.

[13] S. Santini and R. Jain. « Similarity Measures », IEEE Trans. On Pattern Analysis and Machine Intelligence, vol 21, No 9, 871-883, 1999

[14] J. Hérault, « De la rétine biologique aux circuit neuromorphiques », *in Trait. IC2, Les Systèmes de Vision, J.M Jolion ed. Hermès*, 2001.

[15] J. P. Jones, L. A. Palmer, « An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex, *J. Neurophysiol. 58(6):1233-1258*, 1987.

[16] R. L. De Valois & K. K. De Valois, « Spatial Vision ». *Oxford Univ. Press,*1988.

[17] D. McG. Squire, T. Pun, « Assessing Agreement Between Human and Machine Clustering of Image Databases », Patern Recognition, 31, 12, pp 1905-1919, 1998.