Learning Mid-level Image Features for Natural Scene and Texture Classification*

Hervé Le Borgne[1], Anne Guérin-Dugué[2], and Noel E. O'Connor[3]

[1]CEA List, Paris, France (e-mail: herve.le-borgne@cea.fr)
[2]Laboratory of Images and Signals, Grenoble, France
[3]Centre for Digital Video Processing, Dublin City University, Dublin, Ireland

Abstract

This paper deals with coding of natural scenes in order to extract semantic information. We present a new scheme to project natural scenes onto a basis in which each dimension encodes statistically independent information. Basis extraction is performed by Independent Component Analysis (ICA) applied to image patches culled from natural scenes. The study of the resulting coding units (coding filters) extracted from well-chosen categories of images shows that they adapt and respond selectively to discriminant features in natural scenes. Given this basis, we define global and local image signatures relying on the maximal activity of filters on the input image. Locally the construction of the signature takes into account the spatial distribution of the maximal responses within the image. We propose a criterion to reduce the size of the space of representation for faster computation. The proposed approach is tested in the context of texture classification (111 classes), as well as natural scenes classification (11 categories, 2037 images). Using a common protocol, the other commonly used descriptors have at most 47.7% accuracy on average while our method obtains performances of up to 63.8%. We show that this advantage does not depend on the size of the signature and demonstrate the efficiency of the proposed criterion to select ICA filters and reduce the dimension.

1 Introduction

The efficient access and retrieval of visual information from large databases has emerged as a crucial field of research given the increasing number of digital visual documents available, for instance on the Web or in personal and professional picture collections. It has led to the emergence of a new discipline entitled Content-Based Retrieval (CBR), often termed Content-Based Image Retrieval, Content-Based Video Retrieval and more generally Content-Based Multimedia Retrieval (note that although we focus on the visual aspect of the problem in this article, we choose to use the neutral denomination for the sake of simplicity). CBR borrows tools and algorithms from related fields such as pattern recognition, data mining, computer vision and cognitive sciences. One of the key issues to be addressed in CBR is the *semantic gap* – the difference between an image as a mental representation of the visual perception of a human, and a digital image considered as a set of pixels by a computer.

The first step of any CBR system consists of extracting knowledge from the media i.e. the task of feature extraction. Nowadays, it is usual to distinguish between low-level and high-level features. The former refer to primitive features such as color, texture, shape and motion that are derived from the raw pixel values but that *do not refer to any external knowledge* [1]. A major contribution to the definition of these kinds of descriptors was realized during the development of the ISO/MPEG-7 standard. We refer to [2] for a comprehensive presentation of the standard and the corresponding descriptors. High level features, also known as semantic features, on the other hand, generally require human annotation of images (or regions resulting from a segmentation thereof). From this high level expertise, systems can be built to infer automatic annotations for a larger number of images. This class of approaches includes automatic keyword association to images based on converting keyword annotations to a vector containing their frequencies [3] or modelling the joint distribution of images (or regions) and keywords [4]. One can also annotate a test image by comparing it to a learning database and selecting the keywords that are the closest according to a learning framework. In this vein, [5] used a statistical approach and [6] defined an approximative linear discriminant analysis to match words and pictures. [7] used a monotonic tree to cluster low level features and map these last to some keywords to annotate automatically the images. Other approaches consist of including low-level features into an object-oriented database [8], mapping low-level features to high level features using a factor graph framework [9] or enriching an ad-hoc ontology with low-level features [10].

Within the knowledge discovery community, it is becoming clear that both low and high level features must be integrated in a common framework, although the

*Final version published in IEEE T. Circuits and Systems for Video Technologies (2007) http://doi.org/10.1109/TCSVT.2007.890635

practicalities of how this is achieved vary significantly from one work to another. However, it is more or less accepted that low-level features do not carry any semantic information and that they are useful only to enrich a manual annotation.

In this article, we argue that one way to bridge the semantic gap is to define and use low-level features that actually carry, if not semantic knowledge, at least some sense of what is depicted in an image. This behoves us to carefully consider the definition of these features, in particular the fact that they do not refer to any external knowledge of the image database. We propose two contributions in this paper that directly address this issue. First, we propose to learn the features directly from data. Secondly, in order to relate knowledge extraction to human expertise, we propose to drive this feature extraction process on the basis of accepted principles of visual perception. The main contribution of this work is the study of the capacity of Independent Component Analysis (ICA) filters to adapt to the discriminative features of images as well as the proposal of a new type of representation of images using these filters to take advantage of this property. To characterize the discrimination power of ICA filters, we use a model based on a Gabor representation, allowing a description with four parameters. The study of the relationships between couples of parameters shows that the ICA filters adapt to the statistics of the image categories. Building upon this, we proposed an image signature that exploits the discrimination properties emphasised in the former point. Using a support vector classifier, we show that the proposed signature leads to an efficient classification framework that outperforms approaches using other state-of-the-art descriptors.

Learning features directly from data has already been used, perhaps most notably in the definition of eigenfaces [11] to detect and recognize faces. These are defined as the eigenvectors of the scatter matrix of a set of previously normalized faces. This approach has been further adapted and applied to other problems such as texture rendering [12] or 3D object recognition [13]. It is equivalent to the use of Principal Component Analysis (PCA) to provide a new basis of representation in which data is uncorrelated. This strategy can be extended to a biologically inspired approach to image classification and retrieval. Indeed, the formation of the human visual system has been structured by its natural environment through evolution. It has learnt the intrinsic structures of images of the real world and adapted to react to the important salient features of these scenes. This adaptation has been achieved through specific mechanisms and we argue in this work that learning features directly from data by simulating these mechanisms leads to analogous visual detectors well adapted to the discriminative properties of images. As a result of this strategy, the resulting descriptors carry some sense of the image data analyzed.

In his seminal book, Maar proposed three levels to model perception as an information-processing system [14] corresponding to the level of computational theory, the level of the algorithm, and the level of the physical implementation. The first defines the goal of the processing and thus answers the question: *why are the considered inputs transformed into the desired outputs?* The second level considers algorithmic principles, that is to say the coding of the inputs and the way in which they are transformed into the outputs. The third level checks whether the first two can actually be implemented considering neurophysiological constraints. This last level is not considered in our work. The first of Maar's questions is answered in our work by ensuring that during the first steps of visual processing, the inputs are encoded in a non-redundant manner [15]. Such a factorial code is ideally obtained when the coding channels are statistically independent. To achieve this, and thus answer the second of Maar's questions, we use independent component analysis [16, 17] that provides a new basis of representation on which the data is statistically independent.

The remainder of this paper is structured as follows. Section 2 describes the theory of independent component analysis as well as how it has been applied to images in previous works. In section 3, we study the properties of ICA filters extracted from data, in particular we show how they are adapted to the discriminative features of image categories. Section 4 presents the proposed method to describe and classify natural images and textures. In section 5, experimental results using the proposed method are presented and compared with other state of the art techniques. Finally, discussion of the whole work as well as concluding remarks are reported in section 6.

2 Representation of images with independent features

2.1 Independent Component Analysis

Independent component analysis (ICA) is a concept that initially emerged from research in neuroscience for modelling the biological problem of motion coding [18]. It has become popular thanks to its ability to solve the blind source separation (BSS) problem [19] that corresponds to recovering independent sources given only mixtures of these sources (sensor observations). The adjective *blind* simply refers to the fact that both the sources and the mixing function are unknown. Thus, N observations, modelled as an N-dimensional random vector X_N, are assumed to be a linear mixture of M mutually statistically independent sources S_M:

$$X_N = AS_M \tag{1}$$

where A represents a linear mixture called the *mixing matrix*. To achieve the separation, one must estimate the *separating matrix* W that verifies:

$$Y_M = WX_N \tag{2}$$

where Y_M is an estimation of the M sources S_M, and the (pseudo) inverse of matrix W is an estimation of the matrix A. Since both sources S_M and observations X_N are unknown, this is an atypical inverse problem for which classical identification methods cannot be used. However, assuming statistical independence between sources in the model (1), a class of methods that exploit higher order statistics was derived to estimate W and Y_M.

Assuming a linear mixture of independent sources without noise, Comon showed that the ICA/BSS problem is solvable (i.e one can theoretically recover the sources or, by equivalence, the mixture) when at most one Gaussian source is present and the rank of A is equal to the number of sources (i.e there are as many sources as observations: $M = N$) [16]. Several methods were proposed to perform such an estimation, such as minimizing the mutual information between the components [20], approximating with cumulants of increasing order [16], or maximizing the output entropy of a neural network of nonlinear units [21] (i.e information maximization between X and S), which is equivalent to a maximum likelihood approach [22].

In [23], the authors remark that the sum of independent random variables is closer to a Gaussian distribution than any of the independent variables themselves (central limit theorem). Hence, independence between estimates of the sources is achieved by forcing them toward a maximum value of non-Gaussianity. They introduced approximations of the negentropy and derived a fixed-point iteration to estimate the sources. It resulted in the *fast-ICA* algorithm that is used in our work. It converges at least quadratically, while other algorithms based on a gradient descent converge linearly.

Nevertheless, even if the two conditions of identification hold, two ambiguities remain regarding the estimates. First, any permutation on the index of the sources will not change their mutual information thus, contrary to PCA for instance, the sources are not ordered. The second ambiguity relates to the magnitude of the sources that is known give or take a scale factor. In particular, a negative scale factor inverts the sign of the signals.

Within the last ten years, the model (1) has been widely used in diverse areas, such as audio separation, biomedical imaging, analysing financial data and unmixing hyperspectral data. Many references on ICA applications can be found in [17].

2.2 Natural image representation using ICA

The model (1) can be applied to the grey-scale values (point luminance) of natural images. In practice, for computational feasibility, it is applied to small image patches $P(x, y)$. Each image patch is considered as a linear superposition of some basis function $a_i(x, y)$, weighted by some underlying "causes" s_i. Each patch is then represented by a particular sample of these sources, that corresponds to their activities projected on an internal model formed by the basis functions :

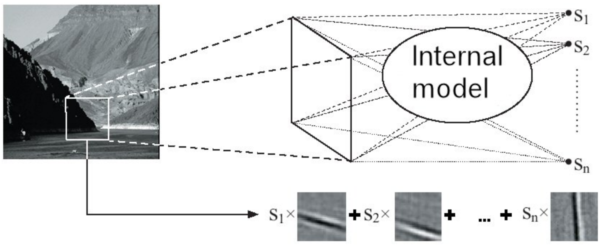

Figure 1: An image (or a part of an image) is modelled as a weighting sum of basis functions (from [52])

$$P(x, y) = \sum_{i=1}^{n} s_i a_i(x, y) \qquad (3)$$

Estimation of this model consists of determining simultaneously $a_i(x, y)$ (consequently $w_i(x, y)$) and the s_i, by exploiting the statistical properties of the inputs.

Olshausen and Fields conjectured that a low-entropy coding of natural images could be found by imposing optimal reconstruction of inputs (minimal mean square error) under sparsity constraints [24]. They obtained a collection of localized and oriented basis functions similar to the simple cells of the visual cortex. Similar results were obtained with other unsupervised learning algorithms [25]. On the other hand, Nadal and Parga in [26] showed the equivalence between the redundancy reduction principle [15] and the infomax principle [27]. Hence, Bell and Sejnowski used ICA [21] as the algorithm level to implement the same conceptual level. This led to similar basis functions to those found by Olshausen and Fields [28]. Van Hateren and Van der Schaaf, using the FastICA algorithm, have shown that most of the properties of these basis functions match well the properties of the receptive fields of simple cells in the cortex of a macaque monkey [29].

In this context, ICA provides not only an estimation of the basis functions of the generative model (1) but also some ICA filters $W(x, y)$ that can be used to analyze natural images. An image $I(x, y)$ is filtered by D_t ICA filters $w_1(x, y), \ldots, w_{D_t}(x, y)$. It is then represented by a multidimensional density $Y\left(I(x, y)\right)$:

$$\begin{aligned} Y\left(I(x,y)\right) &= \left(Y_1\left(I(x,y)\right), \ldots, Y_{D_t}\left(I(x,y)\right)\right) \\ &= \left(I(x,y) \otimes w_1(x,y), \ldots, I(x,y) \otimes w_{D_t}(x,y)\right) \end{aligned} \qquad (4)$$

where $\otimes$ represents the convolution product. Several degrees of complexity were proposed to model the densities of the responses. In [30] the authors use the mean of the density (average energy of the global response) as a signature. They show the validity of the approach by discriminating images of faces, leaves and buildings, objects [31], and natural scenes [32]. In [33], a restricted set of the responses are modelled by Gaussian mixtures allowing invariance to partial occlusion for object recognition. In [34], the marginal densities are estimated by a simple histogram. The sufficiency of this representation is demonstrated in the context of object, texture and face recognition. This model was used in [35] to represent natural image categories. In [36], several signature models are discussed and compared to an unsupervised estimation of the densities. The use of the Kullback-Leibler divergence to compute the distance between densities leads to a synthetic representation scheme for this model. For instance, using the Euclidean distance, the average responses of ICA filters is computationally equivalent to the Kullback-Leibler divergence between Gaussian distributions of common variance for which the means are estimated by the average of ICA filter responses.

2.3 Practical extraction of ICA filters

Let us consider a small set of grey-level natural images. First of all, the luminance of each image is filtered by a non-linear filter that simulates the processing of the retina [37]. It flattens the spectra of the image by enhancing the higher frequencies. Then, a collection of N_{ptch} patches of size $S_p \times S_p$ is extracted at random locations within the images and stored into the columns of matrix X in equation (1) or (2). The data is centred (zero mean) then transformed so that the components are uncorrelated and have unit variance. This is achieved via a principal component analysis (PCA) that is also used to reduce the dimensionality. For this, we compute the eigensystem of the correlation matrix of X, and the data is projected onto the N_{PCA} first eigenvectors. This number is chosen as a compromise between the portion $\frac{N_{PCA}}{S_p^2}$ of the variance retained to encode the data and the computational cost for the ICA estimation. Finally, N_{ICA} filters are iteratively estimated by the fast-ICA algorithm ([23] and section 2.1) using the *tanh* non-linearity, and stored in the matrix W (size $S_p^2 \times N_{ICA}$).

3 Adaptation of filters to image categories

3.1 Gabor parameterization of ICA filters

Most of the ICA filters are localized and oriented band-pass filters. Hence, they can be modelled as Gabor filters or wavelets as a first approximation. Such a model is entirely determined by four parameters that give the position and the shape of the Gaussian envelope in the frequency domain:

$$G(u, v|F_0, \theta_0, \sigma_u, \sigma_v) = exp\left(-\frac{1}{2}\left(\frac{(u - F_0)^2}{\sigma_u^2} - \frac{v^2}{\sigma_v^2}\right)\right) \tag{5}$$

where (F_0, θ_0) are the polar coordinates of the central frequency, and (σ_u, σ_v) are the standard deviations of the Gaussian envelope. The shape factor is defined as $S_f = \frac{\sigma_u}{\sigma_v}$. Hence, a value $S_f = 1$ corresponds to an anisotropic filter. When $S_f < 1$ the filter is stretched along its main direction θ_0, thus more selective in this direction. On the contrary, for $S_f > 1$, the filters are selective along a direction perpendicular to θ_0.

We search for the best Gabor approximation of an ICA filter $F_{ICA}(u, v)$ in the frequency domain, by minimizing the following quadratic criteria:

$$Q = \iint\limits_{\substack{-0.5 \leq u \leq 0.5 \\ -0.5 \leq v \leq 0.5}} \left[\frac{F_{ICA}(u, v)}{max(F_{ICA}(u, v))} - G(u, v|F_0, \theta_0, \sigma_u, \sigma_v)\right]^2 \mathrm{d}u\mathrm{d}v \tag{6}$$

Normalization of the ICA filters by their maximum values does not affect the model (2) because of the ambiguity of their magnitude. Initial values for (F_0, θ_0) are fixed such that each value matches the maximal value of the ICA filter. Then, minimization is performed by conjugate gradient descent, constraining (F_0, θ_0) in the neighbourhood of their initial values and standard deviations between 10^{-4} and 0.3. Other strategies without constraints on the parameters or different normalization of ICA filters were tried, without any qualitative change in the results [38].

3.2 Coupled-parameterization of filters

In their seminal study [29], Van Hateren and Van der Schaaf characterized ICA filters in order to compare their properties to those of the receptive fields of simple cells in macaque monkey cortices. For this, they compared the occurrences of several parameters in both cases. In this paper, we aim to characterize ICA filters in terms of discrimination, which is best achieved by studying the relationship between couples of parameters. Another difference in our approach is the set of images from which the filters are extracted. In [29], the images were chosen as representative inputs of the macaque visual system, capable of influencing its evolution. They thus represented images of natural landscapes in various situations. On the contrary, we want to show here how ICA filters have some spectral properties that are adapted to the discriminative characteristics of the categories they are extracted from. For this reason, we extracted collections of filters from categories of images that are coherent in terms of spectral properties and visual coherency.

We extracted PCA and ICA filters from these categories according to the method previously described (section 2.3) with the parameters $N_{ptch} = 40000$, $S_p = 16$, $N_{PCA} = 150$ and $N_{ICA} = 50$. The average spectrum of each category was computed from N_{ptch} patches (Figure 2, column 2). We also computed the average spectrum of ICA and PCA filters extracted from them (Figure 2, column 3 and 4). From one category to another, the PCA filters have quite a similar average spectrum. On the contrary, the ICA filters adapt differently to each collection of images. As a consequence, PCA filters will produce similar responses to images from one category to another. ICA filters have a varied behaviour as a function of the image category that will ultimately lead to a higher discrimination power. Fundamentally, we interpret this as being due to the relative invariance of the natural image statistics up to the second order [39], while their properties differentiate themselves at a higher order [40, 41].

Each ICA filter has been modelled according to the method described in section 3.1. It resulted in the estimation of the central frequency and the shape factor of each filter. In Figure 3, the central frequency F_0 of each filter is superposed on the average spectrum of the category in the Fourier plan, while the shape factor S_f is plotted according to the direction θ_0 of the filter. One can see that the shape factor is below 1 at the direction 0° and 90° and has a higher value at other directions. This shows how the ICA filters tend to be more selective on the vertical and horizontal directions, which carry not only the largest part of the energy, but also the main differences of the average spectrum from one category to another (Figure 2, column 2).

In summary, we showed that ICA filters adapt to the most discriminative features of the image categories. In the following we exploit this property to define an image signature amenable to efficient classification of natural images.

4 Proposed representation for natural images

The representation of natural images proposed here (see algorithm 1) directly benefits from the selective adaptation of ICA filters to the image categories of the learning database. In this section, we assume a particular set of images considered as a coherent visual category, which is divided into a learning and a testing set. The images are first converted to the YCbCr colorspace that is considered here as an acceptable model of the color-opponent mechanism used in the human retina. This results in one achromatic channel (luminance) and two chromatic channels, named Cb (blue-yellow opposition) and Cr (red-green opposition). The color information is processed separately to the luminance as explained in section 4.2.

4.1 Luminance signature

The signature of a test image $I(x,y)$ for the luminance component is computed according to algorithm 1. D_t ICA filters are extracted from the learning databases, and the convolution of these filters with the images give D_t responses $R_d(x,y)$ of the same size as $I(x,y)$ (only the valid part of the convolution is kept). The map of activity $A_I(x,y)$ of the image contains the index of the most active filter at each pixel. Because of the ambiguity in the sign of filters extracted by ICA, this maximal activity is computed using the absolute value. The global luminance signature is then the histogram of $A_I(x,y)$[1]. This does not take into account the spa-

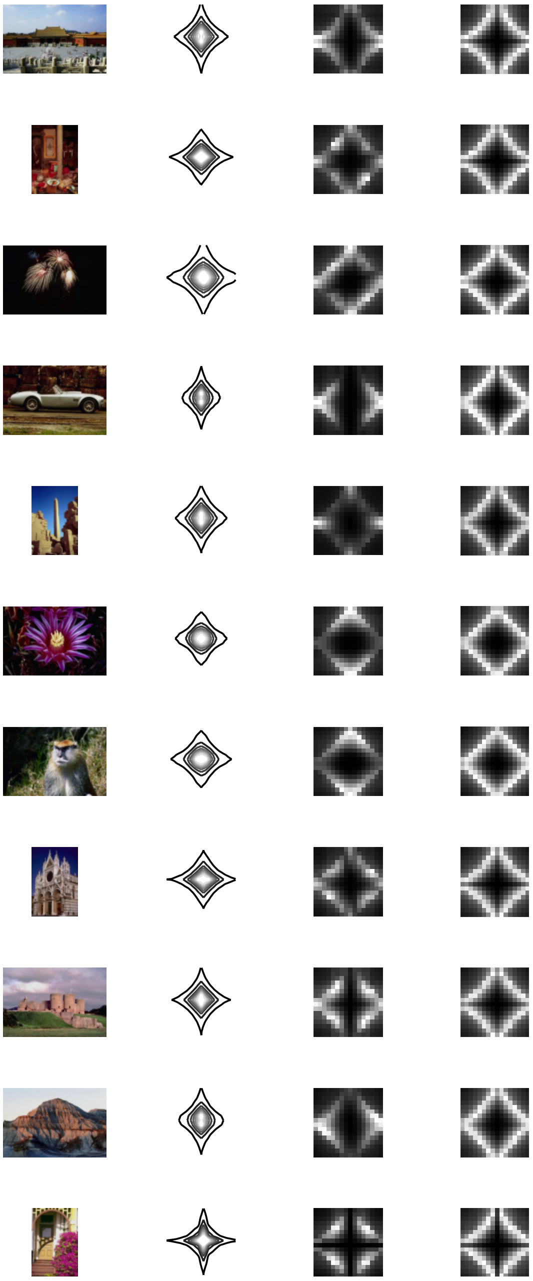

Figure 2: Some examples of image categories (a visual example in column 1; names are given in table 1) with their average spectrum (column 2) as well the average spectrum of ICA (column 3) and PCA (column 4) filters extracted from them

[1] This has already been proposed by A. Labbi in an unpublished report [42]

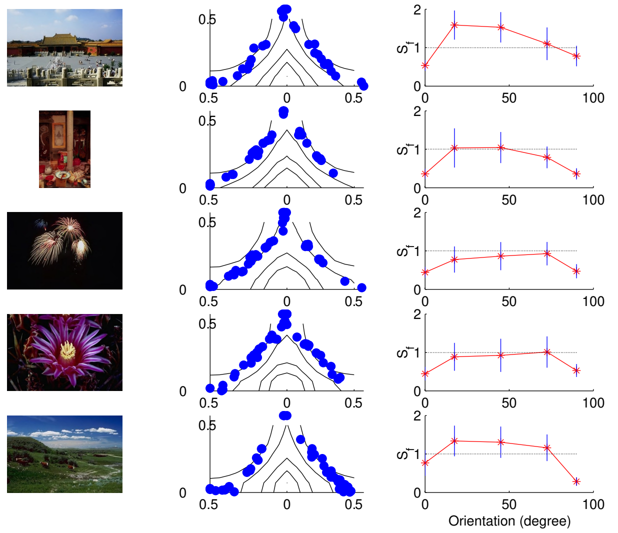

Figure 3: Column 1 : a visual example of some image categories; Column 2 : average spectrum of the category (black lines) and the central frequency of the ICA filters extracted from the category (blue dots); column 3 : shape factor (S_f) of the ICA filters according to their orientation. Five groups of orientation are considered ([0°, 5°],[5°, 45°], [30°, 60°], [45°, 85°], [85°, 90°]) and the mean (± standard deviation) is displayed for each group.

```
Extract ICA filters $F_d$ ($d \in [1, D_t]$)
Let consider one test image $I(x, y)$ (size $S_1 \times S_2$)
Initialize the size of the local structuring element as $s$
for $d \in [1, D_t]$ do
    $R_d(x, y) \leftarrow F_d * I(x, y)$
for $(x, y) \in [1, S_1] \times [1, S_2]$ do
    $A_I(x, y) = arg\max_d (R_d(x, y))$
Initialize $G_I$ as a vector of 0 (length $D_t$)
for ($d \in [1, D_t]$ do
    $G_I(d) \leftarrow Card(A_I(x, y) = d)$
Initialize $L_I$ as a vector of 0 (length $D_t$)
for $(x, y) \in [1, S_1 - s + 1] \times [1, S_2 - s + 1]$ do
    for $(i, j) \in [0, s - 1] \times [0, s - 1]]$ do
        $a(i, j) \leftarrow A(x + i, y + j)$
    $L_I(d) \leftarrow L_I(d) + Card(a(i, j) = d)$
```

Algorithm 1: Computation of the global (G_I) and local (L_I) signatures using ICA filters on a test image (I). By default, ($s = 8$)

tial relationships between pixels but makes sense for the global characterization of a scene that is perceived holistically at first sight [43]. However, at a local scale natural images exhibit meaningful spatial structures that also carry important information [41]. To catch this information, we use a $s \times s$ sliding window $a(i, j)$ that counts the number of times one filter is the most active within a local area of $A_I(x, y)$.

4.2 Color signature and normalisation

For each image, the mean and standard deviation of each color channel is computed. These four features are then merged to the luminance signature. However, because of the large numerical difference, the color part is linearly scaled. The scaling coefficient K_{col} is determined by learning, using an independent validation database.

4.3 Classification scheme

Several classifiers can be used to learn the categories from the produced features. The purpose of this work is mainly to show the interest of learning some features directly from data and the comparison to the state-of-the art will be conducted at this level. As a consequence, the choice of the classifier is not a crucial aspect of our work and we chose to demonstrate our method using a support vector machine (SVM).

Support vector classifiers [44] are commonly used because of several attractive features, such as simplicity of implementation, a small number of free parameters to be tuned, the ability to deal with high-dimensional input data and good generalisation performance on many pattern recognition problems. This last property is due to the fact that this classifier tends to minimise an upper bound on the expected risk (structural risk minimisation), while other learning techniques such as neural networks usually tend to minimise the error on the training set (empirical risk minimisation). To apply a support vector machine to classification in a linear separable case, one considers a set of training samples $\{(x_i, y_i), \ x_i \in \mathcal{X}, \ y_i \in \mathcal{Y}\}$, with $\mathcal{X}$ the input space, and $\mathcal{Y} \triangleq \{-1, +1\}$ the label space. In the linear case, one assumes the existence of a separating hyperplane between the two classes, i.e a function $h(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$ parameterized by $(\boldsymbol{w}, b)$, such that the sign of this function when applied to x_i gives its label. By fixing $\min_i |h(x_i)| = 1$, the normal vector $\boldsymbol{w}$ is fully determined such that the distance from the closest point of the learning set to the hyperplane is $1/\|w\|$. When training data is not linearly separable, a more complex function can be used to describe the boundary. This is done by using a kernel to map non-linear data into a much higher dimensional feature space, in which a simple classification is easier to find.

To classify several classes at the same time, we used a one-against-one strategy. For C classes it consists of training all the $\frac{C(C-1)}{2}$ possible 2-classes classifiers. A given test vector x_t is thus classified between 0 and C times to each category. A majority vote determines the winning class.

4.4 Feature selection

If one set of filters is extracted for each category, the dimension of the luminance signature grows with the number of categories considered. As a consequence, it is desirable to select the dimension (*i.e* the number of features/ICA filters) to reduce this dimension for faster computation. An optimal feature selection for supervised classification requires an exhaustive search that is computationally intractable. A less greedy strategy is to define a criterion to sort the filters, then retain only the N_f first, considered as those leading to the best possible classification rate.

We chose to derive such a criterion from the *dispersal factor* that we previously presented in [45]. The idea is to consider the most useful filters for classification as those providing the most varied responses to a learning database. Indeed, it seems reasonable to think that, inversely, a filter producing similar responses to all images poorly discriminates between categories. The criterion for an ICA filter F_d is computed as follow. Let us consider an image $I_k(x, y)$ of size $S_1 \times S_2$ and its response to the filter $R_d(x, y) = F_d * I_k(x, y)$ ($*$ is the convolution operator). Let consider the average filter response:

$$\overline{R_d(I_k)} = \frac{1}{S_1 S_2}\sum_{x=1}^{S_1}\sum_{y=1}^{S_2}|R_d(x,y)| \tag{7}$$

and standard deviation of the response :

$$\widetilde{R_d(I_k)} = \sqrt{\frac{1}{S_1 S_2 - 1}\sum_{x=1}^{S_1}\sum_{y=1}^{S_2}\left(|R_d(x,y)| - \overline{R_d(I_k)}\right)^2}$$

The dispersal factor is simply the standard deviation of $\overline{R_d(I_k)}$ for all images of a learning database $I_1, \ldots, I_{N_l}$. We define the criterion as the product of the dispersal factor and the average of $\widetilde{R_d(I_k)}$:

$$\zeta(F_d) = \frac{1}{N_l}\sum_{k=1}^{N_l}\widetilde{R_d(I_k)} \times \sqrt{\frac{1}{N_l - 1}\sum_{k=1}^{N_l}\left(\overline{R_d(I_k)} - \frac{1}{N_l}\sum_{k=1}^{N_l}\overline{R_d(I_k)}\right)^2}$$

That can be expressed simpler as :

$$\zeta(F_d) = Avg\left(\widetilde{R_d(I_k)}\right) \times Std\left(\overline{R_d(I_k)}\right) \tag{8}$$

5 Experimental evaluation

One of the difficulties in evaluating CBR algorithms is the lack of annotated databases. This is largely due to the dependence of the ground-truth on a particular task. In other words, it is probably impossible to definitively define a unique database that would match the requirements of any user in any situation. To tackle this difficulty, several strategies can be considered for experimentation purposes, all using a manual annotation of the images.

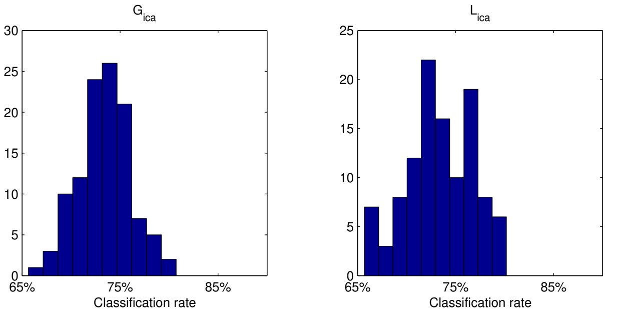

Figure 4: Distribution of the classification rates using the global (left) and local (right) ICA signatures on 111 texture categories.

5.1 Texture classification

Texture is an important feature for image classification. We consider here a set of 111 texture images extracted from the Brodatz album [46]. We derive a category from each image (size 640×640) by dividing it into 25 non-overlapping images of size 128 × 128. The 10 first images are grouped into $Base_1$ (size 1110) and the 15 others into $Base_2$ (size 1665). ICA filters where first extracted on $Base_1$. We extracted ICA filters from each of the categories according to the method previously described (section 2.3) with the parameters $N_{ptch} = 2500$, $S_p = 16$, $N_{PCA} = 150$ and $N_{ICA} = 25$. In practice, it allows to keep more than 95% of the variance in each case.

We run a first experiment by setting $Base_1$ as training set and $Base_2$ as testing set. The classification scheme was the one described in section 4.3. We run 111 experiments with all the images but using only $N_{ICA} = 25$ filters extracted from one category. The results goes from 65.7% to 80.7% for G_{ica} and from 65.7% to 80.2% for L_{ica} (see Figure 4). We compared to the following standard MPEG-7 descriptors [2]: edge histogram (EH), homogeneous texture (HT) as well as the combination of both (EH+HT). We obtained a 58.4% with EH (size 80), 83.5% HT (size 62) and 83.7% with EH+HT (size 142).

We run a second experiment with the same training and testing sets but using several groups of ICA filters. The choice of the groups was done according to their individual performance in the former experiment. In other words we first classified using the 25 filters extracted from the category giving the best results, then 50 filters extracted from the categories giving the two best results and so on. For comparison, we chose the best MPEG-7 descriptors (EH+HT) and run twenty classifications with a random selection of the dimensions, restricted to a specific size ([25, 50, 75, 100, 125, 142]) each time. The average and standard deviation of the twenty classification rates were computed and the minimal and maximal values were collected. All these results are reported on figure 5. It shows that the average classification rates grow with the sizes of the signatures. However, this growth becomes almost null (*i.e* the classification rate is stable) for feature dimensions over 100. Moreover, both

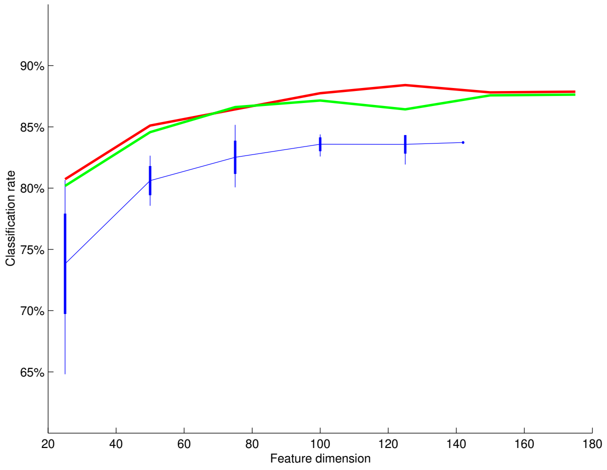

Figure 5: Classification rates according to the feature dimension using the global (red curve) and local (green) ICA signatures and EH+HT (blue) on 111 textures categories. The ICA filters are selected according to their individual performances as reported in Figure 4. The blue curve is the average classification rate for 20 repetitions with a random selection of EH+HT MPEG-7 descriptors. The thick vertical lines show the range at plus or minus one standard deviation. The thin vertical lines show the maximal values.

local and global ICA signatures give better results than MPEG-7 descriptors for all feature dimensions.

We reported in Figure 6 the confusion matrix for the best classification rate (G_{ica} at 125 dimensions). One can see that most of the textures are perfectly classified and that errors are due a very restricted number of confusion. It is the case for instance of textures 50 and 52 as well as 50 and 51 that are represented on the first raw of figure 7. It is also the case for the couples of textures (66, 67), (42, 27) and (36, 103). In most of these cases the confused textures are quite similar.

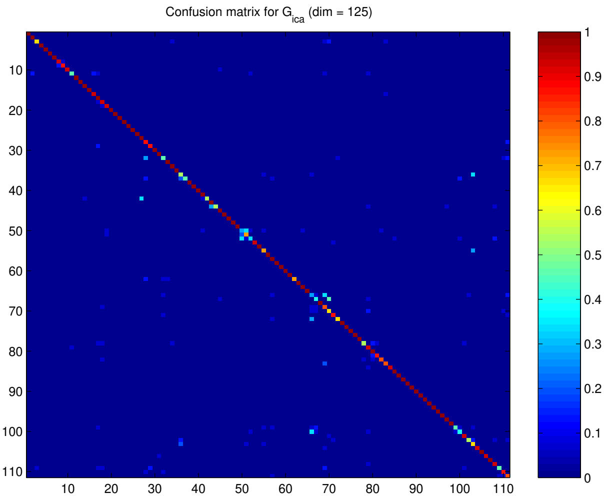

Figure 6: Confusion Matrix for G_{ica} restricted to 125 dimensions.

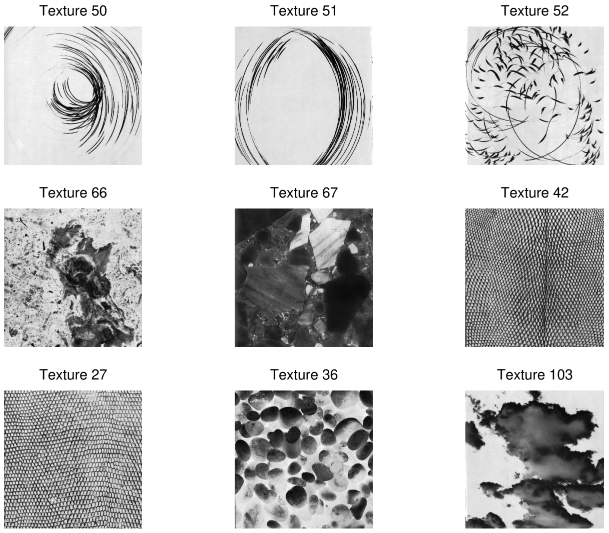

Figure 7: Exemples of various classes of textures.

5.2 Multi-class scene categorization

We describe now an experiment that was conducted on the extensively used COREL database. It consists of a collection of small categories of images, semantically coherent, containing low resolution pictures but of good visual quality, in the sense that the high resolution versions can be used for editorial purposes. We chose 11 categories in such a manner that their annotations correspond to real visual content[2]. The sizes of the different categories are reported in table 1. We extracted ICA filters from these categories according to the method previously described (section 2.3) with the parameters $N_{ptch} = 40000$, $S_p = 16$, $N_{PCA} = 150$ and $N_{ICA} = 50$.

The proposed descriptors are compared to the following standard MPEG-7 descriptors [2]: edge histogram (EH), homogeneous texture (HT), color layout (CL) and scalable color (SC). We consider these descriptors separately as well as combined together. In that last case, they are merged into a unique vector for each image. The classification is achieved using the same learning algorithm as for our descriptors (section 4.3). The support vector classifier is implemented using the LibSVM library [47] with a polynomial kernel of degree 3.

We also compare our approach to some more recent methods based on a SIFT (Scale Invariant Feature Transform) description of images [48]. They are Gaussian derivatives at 8 orientation planes over a 4×4 grid of spatial localizations. They provide a local description at each salient point of the image (points of interest) that is scale invariant. Compared to other local descriptors, SIFT is the best in the context of object classification [49]. From this description, we derive the image signatures using the bag of keypoints (*BoKPts*) technique. First we construct a visual vocabulary (codebook) using K-means on the learning database. Then, for a given test image, we count the number of keypoints associated to each element of the

[2] Some thumbnails of the different categories used in the experiment can be consulted at http://www.eeng.dcu.ie/∼hlborgne/icascene.html

Table 1: Results of the multi-class scene categorization. For each category, the table reports the classification rates with one or several MPEG-7 descriptors and with the ICA signatures (G_{ica} for global, L_{ica} for local)

Class name	Size	EH	CL	SC	HT	EH + CL	EH + SC	HT + CL	HT + SC	EH + CL + SC	G_{ica}	L_{ica}	G_{ica} + color	L_{ICA} + color
Cities	200	33.9	43.3	19.4	25.0	41.7	42.8	36.7	32.8	51.7	**70.0**	46.1	47.8	69.4
Indoor	541	43.8	25.7	15.4	21.1	39.0	48.0	24.2	21.5	43.4	41.1	51.1	55.3	**66.6**
Firework	100	48.8	85.0	73.8	72.5	**96.3**	**96.3**	73.8	73.8	95.0	47.5	91.3	56.3	88.8
Cars	200	45.0	27.2	26.1	22.8	49.4	51.1	33.9	24.4	52.2	44.4	50.0	50.0	**56.7**
Egypt	100	13.8	31.3	**52.5**	10.0	26.3	25.0	15.0	8.8	31.3	25.0	25.0	36.3	33.8
Flowers	400	41.8	24.5	30.0	30.5	39.5	50.8	33.2	32.9	50.3	**82.9**	81.6	70.5	73.2
Monkeys	100	17.5	25.0	23.8	23.8	25.0	26.3	26.3	28.7	32.5	40.0	40.0	**70.0**	62.5
Churches	96	**46.1**	17.1	15.8	42.1	30.3	42.1	43.4	42.1	34.2	35.5	40.8	42.1	36.8
Castles	100	11.3	17.5	20.0	18.8	16.3	20.0	17.5	15.0	16.3	31.3	**50.0**	20.0	27.5
Mountains	100	37.5	32.5	18.8	23.8	28.7	**45.0**	30.0	33.8	38.8	37.5	36.3	**45.0**	43.8
Doors	100	76.3	60.0	33.8	55.0	72.5	65.0	58.8	56.3	68.8	92.5	91.3	80.0	**93.8**
Total	2037	40.1	31.3	25.6	27.9	41.4	47.7	32.4	30.0	47.1	54.0	57.6	55.6	**63.8**

codebook (*i.e* closer to this element according to the euclidean distance). This histogram is then the signature of the image, which is used as input of the SVM [50]. Some authors proposed to use a simpler binary histogram [51] but we found weaker performances on our problem and do not report these results here.

We use 20 images per class for learning and the rest for testing resulting in 220 images for learning and 1817 for testing. The overall classification efficiency of the MPEG-7 descriptors is at most 47.7% when edge histogram is merged with scalable color and less in all other configurations (comprising the use of the four MPEG-7 descriptors considered here that is not reported in table 1). The classification rate increases to 54.0% with the ICA luminance global signature (G_{ica}) and 57.6% with the ICA luminance local signature (L_{ica}). When the color information is added, the results reach 55.6% for the global signature and 63.8% for the local signature. For individual classes, the classification results are often better with the ICA descriptors than the MPEG-7 ones. In particular, there is a counter performance of ICA descriptors for the class "egypt" that can be explained by a larger visual diversity within it. In other words, this class relates strongly to a pure semantical concept, for which the definition of a visually coherent pattern is difficult. The low performance for the class "castles" is explained by a large overlap with the classes "cities" (21.3% for L_{ica} + color) and "churches" (25%). Visually speaking, they can thus be considered as sub-classes of "man-made constructions". The class "churches" is also blended with the class "indoor" (39.5%). However this is relevant since 38 images of this class are indoor views of churches. Over the 30 images of "churches" classified as "indoor", 25 are actually indoor pictures. Among the 5 other images (actually outdoor), 3 were taken at night. When the classes are strongly visually coherent, such as for classes "firework" or "doors", the classifications with ICA signatures lead to very good results, particularly with the local signatures.

The comparison with the Bag-of-Keypoint approach (table 2) shows a weak performance for the Bag-of-Keypoints signature. We tested this feature with different sizes of codebook and obtained at most 26.7% on average, while we can reach twice this score with the ICA signatures. Several reasons can explain this. One could think this feature is not adapted to the type of image that is classified here, since a large part of the work using these feature focussed on object classification (for instance [51]). More likely, the weak performance is due to the small size of the learning database. We used 20 images per category to match the experimental protocol used in the former experiments. Although such a size is sufficient to classify the images using MPEG-7 descriptors or our method, it is not the case for Bag-of-Keypoints. This lack of learning data is particularly noticeable when the size of the codebook is large (1000). In that case, there is a high confusion of all images with the class *firework*. Indeed the images of this class contain less keypoints than the other on average. As a consequence, the signatures of all images are more likely similar to those of this class. Consequently, our method has the advantage to require a much smaller learning database than the Bag-of-Keypoints and perform better in that case.

5.3 Influence of the signature size

One could think these better results for ICA signatures may be simply due to the higher number of descriptors used, although this is not a guarantee of quality as a rule of thumb. For instance, adding homogeneous texture to the three other MPEG-7 descriptors usually lowers the results. However, since 50 filters were computed on each category, the size of the ICA signatures is 750 (754 for colour) while the MPEG-7 signature size is at most 235. To test the influence of this difference, we conducted the following experiment. For all four possible ICA signatures, a random selection of filter was achieved then restricted to a given signature size. For colour signatures, four of

Table 2: Results of the multi-class scene categorization compared to the SIFT-based signatures. For each category, the table reports the classification rates with the Bag of Keypoints, using a codebook of size N_{cb} ($BoK_{N_{cb}}$). Results with ICA signatures (G_{ica} for global, L_{ica} for local) are reported for comparison

Class name	Size	BoK_{50}	BoK_{100}	BoK_{200}	BoK_{1000}	G_{ica}	L_{ica}	G_{ica} + color	L_{ICA} + color
Cities	200	25.0	13.3	11.7	6.7	**70.0**	46.1	47.8	69.4
Indoor	541	21.7	23.8	25.0	6.9	41.1	51.1	55.3	**66.6**
Firework	100	45.0	41.3	37.5	**98.8**	47.5	91.3	56.3	88.8
Cars	200	26.7	36.1	41.7	21.7	44.4	50.0	50.0	**56.7**
Egypt	100	**42.5**	36.3	33.8	28.7	25.0	25.0	36.3	33.8
Flowers	400	20.8	22.6	22.1	3.2	**82.9**	81.6	70.5	73.2
Monkeys	100	15.0	15.0	15.0	2.5	40.0	40.0	**70.0**	62.5
Churches	96	50.0	51.3	51.3	44.7	35.5	40.8	**42.1**	36.8
Castles	100	37.5	36.3	38.8	18.8	31.3	**50.0**	20.0	27.5
Mountains	100	20.0	13.8	17.5	1.3	37.5	36.3	**45.0**	43.8
Doors	100	42.5	41.3	38.8	25.0	92.5	91.3	80.0	**93.8**
Total	2037	26.7	26.7	27.2	15.0	54.0	57.6	55.6	**63.8**

these dimensions were replaced by the colour descriptions. We then froze 20 images per class for learning and the rest for testing resulting in 220 images for learning and 1817 for testing. Then twenty classification iterations were run using different feature dimensions ($[25, 50, 100, 150, 200, 300, 400, 500, 600, 700, 750]$) each time. The average and standard deviation of the twenty classification rates were computed and the minimal and maximal values were collected. All these results are reported on figure 8. It shows that for all the ICA signatures, the average classification rates grows with the size of the signature. However, this growth is very slow and almost null ($i.e$ the classification rate is stable) for feature dimension more than 200. Moreover, better results can be obtained with fewer dimensions than the maximal one. Even with 25 filters, the results are similar to, or better than, the best combination of MPEG-7 descriptors. Finally, the minimal classification rates for the local ICA signatures are similar (47.4% for L_{ica} and 48.2% for $L_{ica} + color$) to this best MPEG-7 combination.

5.4 Filter selection

Using the same experimental protocol we evaluated the efficiency of our criterion ζ to select filters. This last was computed on the learning database according to equation 8. As shown on figure 9, the obtained classification rate (green curve) is most of the time better than the average random selection (and always better than the results obtained with the MPEG-7 descriptors). For comparison, we also plotted the results obtained using the dispersal factor (dotted red curve) that are quite similar. None of them are "optimal" but both give good classification results, in particular for smaller dimension. Indeed, using our criterion ζ, the best classification rates are often reached around 100 dimensions. Beyond this point, it seems that additive filters tend to alter the results, although maintained at a good level.

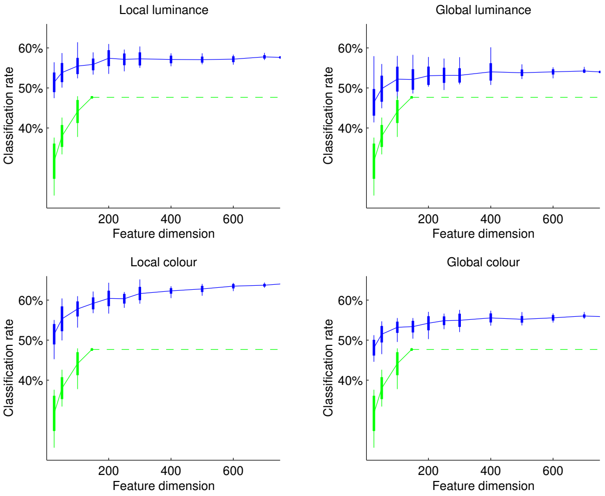

Figure 8: Classification results for the four ICA signatures at different sizes. The dash blue curve is the average classification rate for 20 repetitions with a random selection of ICA filters. The thick vertical lines show the range at plus or minus one standard deviation. The thin vertical lines show the maximal and minimal values. The green curve is the same for the best MPEG-7 classification (EH+SC)

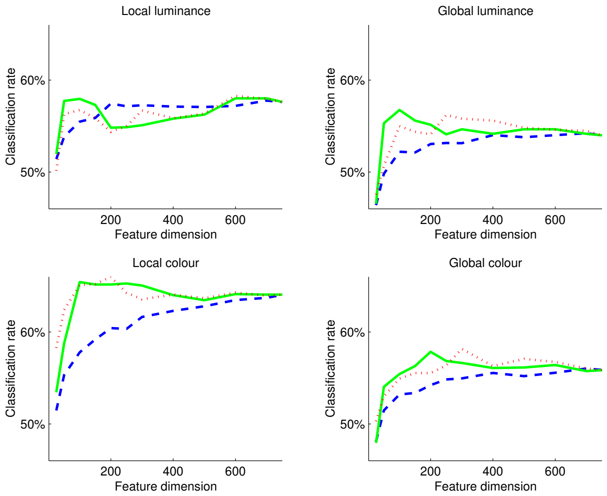

Figure 9: Classification results for the four ICA signatures at different sizes for several methods of filter selection. The dashed blue curve is the average classification rate for 20 repetitions with a random selection of ICA filters (same as the thin blue curve on Figure 8). The dotted red curve is the classification rate when the filters are selected according to their dispersal factor and the thick green curve is the classification rate when the filters are selected according to criterion ζ (Eq 8)

6 Conclusion

We presented a method to learn mid-level features directly from image categories. We used a strategy inspired from visual perception principles postulating that the goal of a vision system is to reduce the information redundancy between the input images and the coded output. To obtain such a code, we used independent component analysis. We showed that taking into account the higher order statistics allows a better adaptation of descriptors to images categories (in comparison with descriptors extracted by principal component analysis that describe them up to the second order statistics). We proposed an algorithm to compute global and local signatures of images using ICA filter collections. They fully take advantage of the properties of adaptation of the filters to the categories, since their definitions rely on the maximal activities of filters applied to natural images and textures. We defined a criterion to select the ICA filters and thus to reduce the dimension of the problem. Combined with a support vector classifier, the proposed signatures lead to an efficient classification framework that outperforms the state of the art descriptors in texture and natural scene classification. We showed this advantage does not depend on the size of the signatures and demonstrated the efficiency of the proposed criterion to select ICA filters. Most of the time, the confusion is due to a very close visual content between the categories.

Since the descriptors are extracted from images, they characterize strongly their visual content. This will lead in future work to their use for giving large image databases a visually coherent organisation. Dealing with such very large databases (10^5 or more images) will require an efficient implementation of our method, adapted to some powerful hardware such as a cluster machine.

Acknowledgment

This work has been partially supported by the European Commission under contract FP6-001765 aceMedia (URL: http://www.acemedia.org). The authors would like to thank the three referees for their valuable comments.

References

[1] C. Djeraba, “Content-based multimedia indexing and retrieval,” *IEEE Multimedia*, vol. 9, no. 2, pp. 18–22, April-June 2002, guest Editor’s introduction.

[2] B. Manjunath, J.-R. Ohm, V. Vasudevan, and A. Yamada, “Color and texture descriptors,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, 2001.

[3] X. S. Zhou and T. S. Huang, “Unifying keywords and visual contents in image retrieval,” *IEEE Multimedia*, vol. 9, no. 2, pp. 23–33, April-June 2002.

[4] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan, “Matching words and pictures,” *Journal of Machine Learning*, vol. 3, pp. 1107–1135, 2003.

[5] J. Li and J. Z. Wang, “Automatic linguistic indexing of pictures by a statistical modeling approach,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 9, pp. 1075–1088, September 2003.

[6] H. Glotin, S. Tollari, and P. Giraudet, “Shape reasoning on mis-segmented and mis-labeled objects using approximated fisher criterion,” *Computer and Graphics*, vol. 30, pp. 177–184, 2006.

[7] Y. Song, W. Wang, and A. Zhang, “Automatic annotation and retrieval of images,” *World Wide Web*, vol. 6, no. 2, pp. 209–231, 2003.

[8] A. Yoshitaka, S. Kishida, M. Hirakawa, and T. Ichikawa, “Knowledge-assisted content based retrieval for multimedia databases,” *IEEE Multimedia*, vol. 1, no. 4, pp. 12–21, April 1994.

[9] M. R. Naphade, I. V. Kozintsev, and T. S. Huang, “Factor graph framework for semantic video indexing,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 1, pp. 40–52, January 2002.

[10] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V.-K. Papastathis, and M. G. Strintzis, “Knowledge-assisted semantic video object detection,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 10, pp. 1210–1224, October 2005.

[11] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of Cognitive Neuroscience*, vol. 3, pp. 71–86, 1991.

[12] K. Nishino, Y. Sato, and K. Ikeuchi, “Eigen-texture method: Appearance compression based on 3d model,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1. IEEE, 23-25 June 1999, pp. 618–624, ft. Collins, CO.

[13] H. Murase and S. K. Nayar, “Visual learning and recognition of 3-d objects from appearance,” *International Journal of Computer Vision*, vol. 14, no. 1, pp. 5–24, 1995.

[14] D. Marr, *Vision.* Freeman publisher, 1982.

[15] H. B. Barlow, “Redundancy reduction revisited,” *Network: computation in Neural Systems*, vol. 12, pp. 241–253, 2001.

[16] P. Comon, “Independent component analysis - a new concept?” *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[17] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* John Wiley and Son, 2001.

[18] J. Hérault, C. Jutten, and B. Ans, “Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé,” in *Actes du Xième colloque GRETSI*, vol. 2, Nice, France, Mai 1985, pp. 1017–1022.

[19] C. Jutten and H. Hérault, “Blind separation of sources, part i: an adaptative algorithm based on neuromimetic architecture,” *Signal Processing*, vol. 24, pp. 1–10, 1991.

[20] A. Amari, A. Cichocki, and H. Yang, “A new learning algorithm for blind signal separation,” in *Advances in neural information processing systems*, M. M. D. Touretzky and M. Hasselmo, Eds., vol. 8. Cambridge MA: MIT press, 1996, pp. 757–763.

[21] A. Bell and T. J. Sejnowsky, “An information-maximization approach to blind separation and blind deconvolution,” *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[22] D. Pham, P. Garat, and C. Jutten, “Separation of a mixture of independent sources through a maximum likelihood approach,” in *Proceedings of EUSIPCO*, 1992, pp. 771–774.

[23] A. Hyvärinen and E. Oja, “A fast fixed-pointalgorithm for independent component analysis,” *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[24] B. Olshausen and D. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, vol. 381, pp. 607–609, 1996.

[25] G. Harpur and R. Prager, “Development of low entropy coding in a recurrent network,” *Network: computation in neural systems*, vol. 7, pp. 277–284, 1996.

[26] J.-P. Nadal and N. Parga, “Non linear neurons in the low noise limit: a factorial code maximises information transfer,” *Natwork: computation in neural systems*, vol. 5, pp. 565–581, 1994.

[27] R. Linsker, “Self-organisation in a perceptual network,” *IEEE Computer*, vol. 21, pp. 105–117, 1988.

[28] A. Bell and T. J. Sejnowsky, “The independent components of natural images are edge filters,” *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[29] J. van Hateren and A. van der Schaaf, “Independent component filters of natural images compared with simple cells in primary visual cortex,” *Proc. of the Royal Society Series B*, pp. 2315–2320, 1998.

[30] A. Labbi, H. Bosch, and C. Pellegrini, “Image categorization using independent component,” in *ACAI workshop on biologically inspired machine learning (BIML’99)*, Crete, Grece, July 1999, invited paper.

[31] ——, “Viewpoint-invariant object recognition using independent component analysis,” in *NOLTA’99*, Hawaï, USA, 28 nov - 3 dec 1999.

[32] ——, “High order statistics for image classification,” *International Journal of Neural Systens*, vol. 11, no. 4, pp. 371–377, 2001.

[33] B. Moghaddam, D. Guillamet, and J. Vitria, “Local appearance-based models using high-order statistics of image features,” in *Conference on Computer Vision and Pattern Recognition (CVPR’03)*, Madison, Wisconsin, 2003.

[34] X. Liu and L. Cheng, “Independent spectral representation of images for recognition,” *Journal of the optical society of america*, vol. 20, no. 7, pp. 1271–1282, july 2003.

[35] J. Lindgren and A. Hyvärinen, “Learning high-level independent components of images through a spectral representation,” in *Proceedings of International Conference on Pattern Recognition (ICPR)*, vol. 2, 2004, pp. 72–75, cambridge, UK.

[36] H. Le Borgne, A. Guérin-Dugué, and A. Antoniadis, “Representation of images for classification with independent features,” *Pattern Recognition Letters*, vol. 25, no. 2, pp. 141–154, jan 2004.

[37] J. Hérault, “De la rétine biologique aux circuits neuromorphiques,” in Les systèmes de vision, ser. IC2, J. Jolion, Ed. Paris: Hermes, 2001, ch. 3.

[38] H. Le Borgne, “Analyse de scènes naturelles par composantes indépendantes,” PhD thesis, Intitut National Polytechnique de Grenoble, Grenoble, France, January 2004.

[39] D. Ruderman, “The statistics of natural images,” *Natwork: computation in neural systems*, vol. 5, pp. 517–548, 1994.

[40] E. P. Simoncelli and B. A. Olshausen, “Natural image statistics and neural representation,” *Annual review of neurosciences*, vol. 24, pp. 1193–1216, 2001.

[41] A. Torralba and A. Oliva, “Statistics of natural images categories,” *Network: Computation in Neural Systems*, vol. 14, pp. 391–412, 2003.

[42] A. Labbi, “Sparse-distributed codes for image categorization,” University of Geneva, Dept. of computer science,” Technical report on ICA and image coding, 1999.

[43] A. Oliva and A. Torralba, “Modeling the shape of the scene: a holistic representation of the spatial envelope,” *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[44] V. Vapnik, *The Nature of Statistical Learning Theory.* NY:Springer-Verlag, 1995.

[45] H. Le Borgne and Guérin-Dugué, “Sparse-dispersed coding and images discrimination with independent component analysis,” in *Third International Conference on Independent Component Analysis and Signal Separation*, December 9-13, 2001, san Diego, CA, USA.

[46] P. Brodatz, *Texture: a photographic album for artists and designers.* Dover, 1996.

[47] C. Chang and C. Lin, *LIBSVM : a library for support vector machines*, 2001, software available at `www.csie.ntu.edu.tw/~cjlin/libsvm`.

[48] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[49] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition.* IEEE, 2003, madison, Wisconsin.

[50] G. Csurka, C. Bray, C. Dance, and L. Fan, “Visual categorization with bags of keypoints,” in *Proceedings of IEEE Workshop on Statistical Learning in Computer Vision, ECCV.* Springer-Verlag LNCS Volumes 3021-3024, 2004, pp. 1–22, prague, Czech Republic.

[51] E. Nowak, F. Jurie, and B. Triggs, “Sampling strategies for bag-of-features image classification,” in *Proc. European Conference on Computer Vision 2006 (ECCV’06).* IEEE, May 7 - 13 2006, graz, Austria.

[52] B. Olshausen and D. Field, “Sparse coding with an overcomplete basis set: A strategy employed by v1?” *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.