Representation of images for classification with independent features

Hervé Le Borgne [a*], Anne Guérin-Dugué [b], Anestis Antoniadis [c]

[a] *Laboratoire des Images et Signaux, Institut National Polytechnique de Grenoble, INPG-LIS, 46 av. Félix Viallet, 38031 Grenoble Cedex, France*

[b] *Communication Langagière et Intéraction Personne Systeme, CLIPS UMR 5524, 385, rue de la Bibliothèque - B.P. 53 - 38041 Grenoble Cedex 9, France*

[c] *Laboratoire de Modélisation et Calcul, IMAG, LMC, BP 53, 38041 Grenoble Cedex 9*

Received 17 December 2002; received in revised form 30 July 2003.
Pattern Recognition Letters (2004) *in press.*

Abstract

In this study, Independent Component Analysis (ICA) is used to compute features extracted from natural images. The use of ICA is justified in the context of classification of natural images for two reasons. On the one hand the model of image suggests that the underlying statistical principles may be the same as those that determine the structure of the visual cortex. As a consequence, the filters that ICA produces are adapted to the statistics of natural images. On the other hand, we adopt a non parametric approach that require density estimation in many dimensions, and independence between features appears as a solution to overthrow the "curse of dimensionality". Hence we introduce several signatures of natural images that use these feature, and we define some similarity measures that correspond to these signatures. These signatures appear as more and more accurate estimations of densities, and the associated distances as estimations of the Kullback-Leibler divergence between the densities. Efficiency of the couple signature/distance is estimated by a K-nearest neighbour classifier, with a "leave-one-out" procedure for all the signatures we define, and a "bootstrap" based one for the best results.

Keywords: Independent Component Analysis, Kullback Leibler divergence, Logspline density estimation, Image distances.

1. Introduction

The growing size of the contemporary digital image libraries has created new needs for the users, like publishers or journalists, criminologists (criminal identification), business people (trademark description), artists and teachers (encyclopaedias), or simply any digital camera owner. As a consequence, it constraints to an automatic indexing, and to directly extract information from images, without any human interpretation. This information is often extracted with "low level features" (colour, texture...) from raw pixel values (Cox *et al*., 1996) and could efficiently discriminate broad classes of images (Guérin-Dugué and Oliva, 2000; Szummer and Picard, 1998; Vailaya *et al*., 1998). These classes of images correspond to semantic groups and could only be defined according to human judgement (Vailaya *et al*., 1998; Rogowitz *et al*., 1998; Guyader *et al*., 2002; Le Borgne *et al*., 2003;). The recent Content Based Image Retrieval (CBIR) systems that were developed are widely based on extraction of low level image features that are stored in multidimensional histograms. See (Johansson 2002; Del Bimbo, 1999) for a complete review of existing systems. Hence, dissimilarity between images are estimated as dissimilarity between multidimensional histograms (Puzicha *et al*., 1999; Sticker & Orengo, 1995), even if it does not match with human judgement of similarity between images.

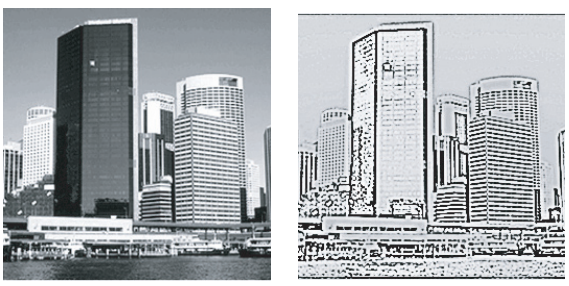
In this article we describe a feature extraction methodology using Independent Component Analysis (Comon, 1994; Hyvärinen *et al*., 2001a) in order to discriminate natural images. When ICA is applied to a set of natural images, it provides band-pass-oriented filters, similar to simple cells of the primary visual cortex (Van Hateren & Van Der Schaaf,

* Corresponding Author.
E-mail addresses : hleborgn@lis.inpg.fr,
anne.guerin@imag.fr,
Anestis.Antoniadis@imag.fr

1998). These filters compose a new basis function set in which images are encoded by independent features. Since it reduces the redundancy between coding units, this model has created great interest, suggesting that the underlying statistical principles may be the same as those that determine the structure of the cortical visual code (Olshausen & Fields, 1997; Bell & Sejnowski, 1997; Labbi *et al*., 1999). Conversely, independent component filters emerge in an unsupervised manner from images and are statistically adapted to these data (Van der Schaaf & Van Hateren, 1996; Le Borgne & Guérin-Dugué, 2001). In this paper we investigate the advantages of this adaptation to the data in context of an image classification task.

Nevertheless "biological plausibility" is not sufficient unto itself to justify the use of ICA for image discrimination. In a given classification problem, it has been shown (Fukunaga, 1990) that the optimal classifier we can design, in the sense it minimises the misclassification risk, is the Bayes classifier. It is equivalent to the maximum *a posteriori* (MAP) classifier which attributes a given vector **x** we want classify to the most probable class. Therefore it requires an estimation of the posterior probability of each class of images, from observations which are known to belong to these classes. Since images are encoded by several features, it follows that the posterior probability is depicted as a multidimensional distribution. We thus discern two approaches, whether we make assumptions or not about the shape of the distribution. In a "parametric approach" we apply some constraints to the distribution and attempt to find the value of the parameters which bring the closest model of data. We can find such a parametric approach in (Do & Vetterli, 2002), where the distributions of wavelet coefficients are modelled with generalised Gaussian densities. (Vailaya *et al*., 2001) have also adopted a parametric approach, with vector quantization. The size of the mixtures (which is also the codebook size for vector quantization) is then a crucial choice which is computationally demanding.

In this paper our philosophy is to remain as less constricting as possible. As a consequence we have chosen a nonparametric approach without any *a priori* over the shape of densities. In that case we are confronted with the well known "curse of dimensionality" which is used to describe the problems associated with the feasibility of density estimation in many dimensions. These problems result from the empirical fact that when the dimensionality of a multidimensional space becomes large, samples quickly become "lost" in this space, and local neighbourhoods become devoid. Indeed, an acceptable estimation of density requires a number of samples that increases more than exponentially when the number of dimension increases. Practically this phenomenon deters correct estimation of probability densities in more than ten dimensions. However we notice that in practice the number of features that encode images can easily exceeds ten (Johansson, 2002), and that in practice the number of samples is limited. In that context, independence between features appears as the only way to make a correct estimation of the desired densities, since in that case (and in that case only), a multidimensional density can always be factorised into the product of the marginal densities. Thus we are in presence of a one-dimensional density estimation problem, which can be solved by classical techniques (Silverman, 1986). These considerations justify the use of ICA to extract a set of basis function in which images are encoded by independent features.



(a) (b)

Fig. 1. (a) A natural image, (b) The same image after whitening.

When natural images are described in terms of a linear superposition of such basis function, they present a "sparse" probability distribution (Olshausen & Fields, 1997). It means the density is highly peaked around zero with heavy tails. In (Hyvärinen *et al*., 2001b) such distributions were modelled by exponential parametric densities like the generalized Laplacian density. In the case of our non parametric approach, we chose the logspline density estimation (Kooperberg & Stone, 1992) which is particularly well adapted to the estimation of exponential families of distribution, since it fits the logarithm of the density we want estimate with "smooth" functions called splines.

The measure of similarity we can associate to the features extracted using ICA directly result from the choice of independence between them. Indeed, independence of a set of random variables is statistically defined as the equality of the joint distribution of the variables and the product of their marginal probability density function (pdf). Hence, the Kullback-Leibler (KL) information, which is precisely defined as a comparison of the true distribution and a statistical model, appears as a natural measure of similarity for our problem. In this paper we use the Kullback-Leibler divergence that we define in part 3.1.

The outline of this paper is as follow. In section 2 we explain the methodology for learning components from images, the pre-processing strategies,

and the way to compute the description of images in the new base of independent features. We also give some details about ICA. Section 3 deals with the signatures we can associate to the images, and the corresponding similarity functions. We also present the "logspline model" for density estimation in detail. In section 4, we present a quantitative appreciation of the efficiency of the signatures we have defined in the prior section, though a classification paradigm. Conclusion and discussions are drawn in section 5.

2. Learning Independent Components from Images

2.1. Database

The training image database from which we extract patches consists on a collection of 540 natural images (256 x 256 pixels, and 256 grey-level values) extracted from several database, and reclaimed on the world wide web. In average, the amplitude spectrum of natural images falls with the spatial radial frequency as $1/f^{\alpha}$, with a fall off factor α between 0.9 and 1.2 (Van der Schaaf and Van Hateren 1996). This factor can be distinguished according to the orientation of spatial frequencies. Considering its variation versus orientation, different shapes of amplitude spectra can be considered corresponding to different semantic categories (Oliva *et al*., 1999). In this study, we consider four categories, containing about 135 images each. Man-made scenes are characterised by horizontal and vertical structures, and include "urban outdoor scenes" and "indoor scenes". Urban scenes contain more horizontal low frequencies (broad vertical edges), while indoor scenes are well balanced between 0° and 90° orientation at all scales. The third category is "open landscapes" (fields, beaches, deserts…) which is characterised by a horizon line, and the fourth category is "closed landscapes" which contains textured scenes without preferential direction (mountains, valleys, forests…). The label of images (their category) was established according to a human judgement (Guyader *et al*., 2002, Le Borgne *et al*., 2003).

A part of these images (4 x 50 = 200) is used to learn "independent components" as we explain in the following paragraphs. The whole set of 540 images is used for the task of classification.

2.2. Image Pre-processing

In a previous paper (Le Borgne & Guérin-Dugué, 2001), we implemented two multiresolution pyramids (3 levels: 256x256, 128x128 and 64x64). The first one was is a low-pass pyramid based on a 6^{th} order low pass Butterworth filter with a cut-off frequency (0.4), and the second one was a band-pass-whitening pyramid. The whitening filter has been implemented according to a biological model of the retina of vertebrates (Hérault, 2001) which realises a nonlinear processing as illustrated on figure 1.

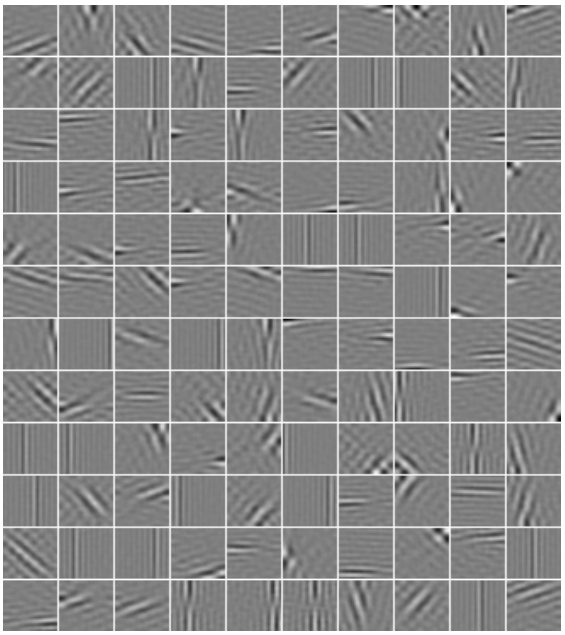


Fig. 2. Exemple of extracted ICA filters.

Here we have only kept the best strategy which was a medium resolution (128x128 pixels) obtained by sub-sampling original 256x256 images, after a low pass filtering (cut-off frequency 0.2) to avoid aliasing, and a band-pass-whitening filtering (Figure 1).

2.3. Principles of Independent Component Analysis

The independent component analysis (ICA) was initially introduced by Hérault, Jutten and Ans in order to perform blind source separation (Hérault *et al*., 1985), but was rigorously defined by (Comon, 1994). In its simplest form, it is an algorithm that search for a linear transform that minimises the statistical dependence between the components of an input vector. Several criterions were proposed to perform such a transform, like minimizing an approximate of the mutual information between the components with cumulants of increasing orders (Comon, 1994), maximising the output entropy of a neural network of nonlinear units (Bell & Sejnowsky, 1995) which is itself equivalent to a maximum likelihood approach (Pham *et al*., 1992). In (Hyvärinen & Oja, 1997), the authors remark that since the sum of independent random variables has a distribution that is closer to Gaussian than any of the independent variables (according to the Central Limit Theorem), they can use measures of non-Gaussianity for ICA estimation. In the same paper they show that this approach is equivalent to a minimum mutual information one (Amari et al., 1996). They introduce approximations of negentropy (which is a modified version of the differential entropy), and derive a fixed-point iteration scheme for ICA estimation. This algorithm is called the "Fast-ICA" algorithm since its convergence is at least quadratic while other ICA algorithms based on gradient descent methods have only a linear conver-

gence.

ICA performs the blind source separation problem with a minimum number of assumptions, since it can estimate the source signals and the function that mixes them, with only one hypothesis of statistical independence between the sources. Nevertheless, two ambiguities remain on the estimates. The first is that their magnitude is known give or take a scale factor. Note that a particular case is a scale factor of -1 which inverts the sign of the signals. The second ambiguity is that, contrary to a principal component analysis for instance, we can not give an order to the components that are estimated, and a permutation of them would not change the result.

When ICA is applied to natural images (Bell & Sejnowski, 1997), it produces sets of visual filters which look like simple cells in primary visual cortex (Van Hateren & Van der Schaaf, 1998), since they can be characterised as being spatially localised, oriented and selective to structure at different spatial scales (Figure 2).

2.4. Estimation of signatures

For each category, we select 50 images from which we extract at random 10,000 patches (at the rate of 200 patches an image) of size 32x32 pixels. In order to minimise the anisotropy on horizontal and vertical orientation, each patch is focused by a weighting Hamming window. Moreover, such a round and smooth window is more biologically plausible (Hurri, 1997). Since it cuts back information all around the patch, the intrinsic dimension is about 700 (instead of 1024). Before ICA, a principal component analysis (PCA) realises data whitening and a dimension reduction from 700 to 225 dimensions. It enables us to retain 92% to 95% of the total inertia.

The model of image we use was proposed in (Olshausen & Field, 1997), and the first who estimate it with ICA were (Bell & Sejnowski, 1997). In this model, we assume that each patch P(x,y) is an independent combination of a set of primitives $\{\phi_i(x,y), i=1..225\}$. The primitives represent the spatial patterns occurring in the different scenes, such as the projection on this basis involves independent codes $\{a_i, i=1..225\}$:

$$P(x,y) = \sum_{i=1}^{225} a_i . \phi_i(x,y) \qquad (1)$$

Practically, we use the "Fast-ICA" algorithm (Hyvärinen & Oja, 1997) with the symmetric method, because of its fast convergence time. It provides four collections of 225 primitives, that we consider as 2D filters $\{F_i(x,y), i=1..225\}$.

We select N filters $\{F_i, i=1...N\}$ according to their "dispersal", *i.e.* the standard deviation of their average response over a collection of image (Willemore *et al.*, 2000). In (Le Borgne & Guérin-Dugué, 2001), we have shown that this criterion is efficient for the selection of ICA filters in the context of an image discrimination task. Each image is thus characterised by a collection of N responses which are considered as particular observations of random variables $\{R_i; i=1...N\}$. The energetic responses of an image I(x,y) to the selected pool of filters are estimated as follow

$$\forall i \in [\![1, N]\!], r_i = \left(I * F_i\right)^2 \qquad (2)$$

They are now considered as the signatures of the image. The squaring operation, which correspond to the energy of the response, results from the intrinsic ambiguity about the sign of the signals that are estimated with ICA, as we explain in part 2.3. One could imagine using the absolute value instead of squaring, and the response model that we present in the next part remain valid. Moreover all the experiments that we present in part 4 were conducted using both signatures. We have chosen the energetic response since we observed better results for classification when we use the energy as a signature, but our comments about the accuracy of the model of response and its consequences on the recognition rate in a classification paradigm remain the same regardless of the signature we use.

We calculate the energetic response of the 540 images to the filters. Images are 128x128 pixels and filters are 32x32, but since we only keep the "valid" part of the response, we dispose $N_k=(128-31)^2=9409$ observations $\{r_i(k); k=1...N_k\}$ of each random variable R_i. The model we choose for these random variables, and especially the quality of the model, determines the distance to calculate differences between images. In the next part, we will consider increasing complex model for the signatures: mean value of responses only, mean and variance, histogram, and finally a model of the whole response.

3. Response model for classification

3.1 Kullback-Leibler divergence

The Kullback-Leibler divergence (or information divergence) is a measure of discrimination between two densities f_1 and f_2. Another name for this measure is the relative (or differential) entropy, and is defined as:

$$KL(f_1, f_2) = -\int_{\mathbb{R}} f_1(x) \log\left\{\frac{f_1(x)}{f_2(x)}\right\} dx \qquad (3)$$

Thanks to the concavity of the logarithm function, this measure is positive when f_1 and f_2 are different, and is zero if f_1 is equal to f_2. Nevertheless, the Kullback-Leibler divergence is neither symmetric nor fulfils the triangle inequality. One of these two drawbacks is solved using a symmetric version of this measure:

$$KL_S(f_1 \,||\, f_2) = KL(f_1, f_2) + KL(f_2, f_1) \qquad (4)$$

If we consider independent variables $\{R_1, R_2,..., R_N\}$ which have densities $\{f_1 f_2,..., f_N\}$, we can factorise their joint probability density functions (pdf) f as:

$$f(x_1,...,x_N)=\prod_{i=1}^{N} f_i(x_i) \qquad (5)$$

Thus the Kullback-Leibler divergence between two multidimensional distributions with independent components is the sum of the Kullback-Leibler divergences between each component.

$$KL(f,g)=\sum_{i=1}^{N} KL(f_i,g_i) \qquad (6)$$

This formula justifies the use of Independent Component Analysis to extract features, since ICA provides filters F_i which analyse an image I in term of independent components r_i (equation 2). Moreover, the Kullback-Leibler divergence between a joint density and the product of the marginal densities is a measure of mutual independence between the corresponding variables, which is minimised in the case of ICA. Thus when data is depicted through an ICA basis functions set, the Kullback-Leibler divergence appears as a natural measure of similarity.

3.2 One or two parameters based model

We first model responses of ICA filters to images (*i.e.* signatures of images) by only one parameter for each dimension (*i.e.* each ICA filter). In that case, the least-squares estimate for this parameter is the mean value, and the distance between the signatures can be estimated with the Euclidean distance.

This point of view is equivalent to consider that the densities are modelled by a Gaussian distribution with same mean as the densities and a common variance. In that case, the Kullback-Leibler divergence applied to Gaussian distribution of same variance is the same as the Euclidean distance of their means (see equation 7 with $\sigma_1=\sigma_2$).

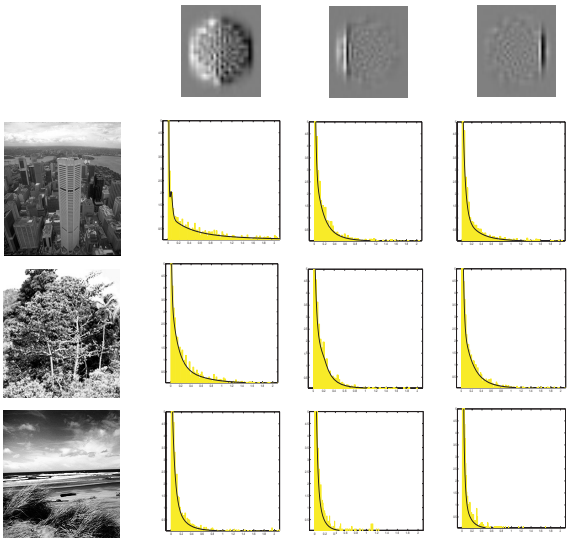


Fig. 3. Exemple of feature for three images (first column) and three filters (first row) - bars = histograms (256 bins of equal length), solid line = logspline density estimate.

Thence, a two-parameters based model can be introduced considering that the signatures are Gaussian distributions, defined by their mean and variance. If a Gaussian density g_1 (respectively g_2) has a mean μ_1 and a variance σ_1 (respectively μ_2 and σ_2), the Kullback-Leibler divergence in its symmetric version is:

$$KL_G(g_1 \| g_2)=\frac{(\sigma_1^2-\sigma_2^2)^2+(\sigma_1^2+\sigma_2^2).(\mu_1-\mu_2)^2}{2.\sigma_1^2.\sigma_2^2} \qquad (7)$$

See (Basseville, 1996) for details of the calculus. It is well worth noting that these models are above all a unified point of view that allows the use of the Kullback-Leibler divergence. The Euclidean distance between μ_1 (which is the mean of a density f_1) and μ_2 (which is the mean of a density f_2) is strictly equal to the KL divergence between a Gaussian density g_1 of mean μ_1 and a Gaussian density g_2 of mean μ_2, with any common variance. Likewise, we will use equation (7) to estimate the distance between f_1 (modelled by its mean μ_1 and its variance σ_1) and f_2 (modelled by its mean μ_2 and its variance σ_2), that is strictly equal to the KL divergence between a Gaussian density g_1 of mean μ_1 and variance σ_1, and a Gaussian density g_2 of mean μ_2 and variance σ_2.

One could be surprised that our two parameters model is equivalent to fit a Gaussian to data that is left censured (equation 2). Thus we introduce an other one-parameter model that consists of fitting a half-normal distribution to data. A half-normal distribution is a normal distribution with mean 0 and standard deviation $1/\theta$ limited to the domain $[0, +\infty)$. In that case the mean of the half-normal distribution (first moment) is $1/\theta$. This value is fitted to the means μ_1 and μ_2 of the responses f_1 and f_2 we want model, and one can deduce the Kullback-Leibler divergence between them from equation (7):

$$KL(f_1 \| f_2)=\frac{\left(\mu_1^2-\mu_2^2\right)^2}{\mu_1^2.\mu_2^2} \qquad (8)$$

3.3 Histogram-based model

We define signatures of images in term of histograms because it provides more complete information about the responses of ICA filters to images.

Let B be the number of bins (we discuss this choice below), V_M the maximum value of all the observations, and N_k the number of available samples. We can compute histogram H using bins H(b) of equal length between 0 and V_M:

$$\forall b \in [1,B],$$
$$H(b)=Card(r_i(k)\cap D_b; k\in[\![1,N_k]\!]) \qquad (9)$$

$$D_b = \left\{ x; \frac{(b-1)*V_M}{B} < x \leq \frac{b*V_M}{B} \right\}$$

Then we normalise its inertia to 1 since we want estimate a density:

$$\forall b \in [1, B], H(b) = \frac{H(b)}{\frac{V_M}{B} * \sum_{b=1}^{B} H(b)} \quad (10)$$

When histograms are the signatures of images, we use the Kullback-Leibler divergence as a measure of dissimilarity between images. For histograms H_1 and H_2 computed with the same number B of bins, it gives:

$$KL_H(H_1, H_2) = \frac{V_M}{B} * \sum_{b=1}^{B} H_1(b) \log \frac{H_1(b)}{H_2(b)} \quad (11)$$

The constant before the sum is the bin-width and then equation (11) corresponds to the rectangular numerical integration.

The choice of the number B of bins, which is equivalent to choose their width according to equation (9), is critical. An efficient, unbiased estimation of the probability density function is achieved when the bin width W is:

$$W = 2*IQR*N_k^{-1/3} \quad (12)$$

Where IQR (interquartile range) is the 75th percentile minus the 25th percentile of the distribution. This result is due to Diaconis and Freedman (Izenman, 1991).

Nevertheless, in practice the responses of images to ICA filters are very sparse (Olshausen & Fields, 1997) (Figure 3), therefore a lot of sample values are close to zero and the interquartile range will be small while the maximum value of the samples could be more than twenty. In these conditions, equation (12) leads to a number of bins that can reach several hundreds. In reason of the finite size of images, we have a limited number of samples (9409 samples for images of size 128x128), then such histograms will poorly estimate some parts of the densities. That is why in practice one solution is to estimate the interquartile range on the logarithm of data. An other solution is to renounce to a constant bin width, and adopt a logarithmic scale:

$$D_b = \left\{ \begin{array}{l} x; 10^{\chi + \frac{(b-1)*(\log_{10}(V_M) - \chi)}{B}} < x \\ x; x \leq 10^{\chi + \frac{b*(\log_{10}(V_M) - \chi)}{B}} \end{array} \right\} \quad (13)$$

where χ is the base ten logarithm of the floating point relative accuracy of the machine on which we calculate histograms. In other words, ten to the power χ is the smallest value we can calculate. Compute an histogram according equation (13) is the same as estimating the density of the logarithm of data. The histogram is then normalised according to the support and the distance is estimated according to Eq. (11).

3.4 Logspline model

3.4.1. Logspline densities based on B-spline

The most complete information we can obtain from the responses of ICA filters to images is contained in the (true) density function of these responses. It exists several methods of estimating an unknown density function from sample data. Histograms are simple estimates of these densities, but strongly depend on choices about the number of bins and their distribution. The most popular methods are kernel based methods, which are well studied in (Silverman, 1986). If we have N sample data $y_1, ..., y_N$, then the estimator of the density function has the form:

$$\hat{f}(y) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{w_i} K\left(\frac{y - y_i}{w_i} \right), \quad y \in \mathbb{R} \quad (14)$$

where K is a Gaussian kernel and w_i is the width of the kernel. When we construct such an estimate, the choice of the widths is critical. If we chose them too small, we take the risk of introducing features that are not really significant. If we choose them too large, we risk to lose important parts of the density that are crucial for discrimination.

Logspline density estimation (Kooperberg and Stone 1992) is an automated methodology for using cubic splines with linear tail in order to model the logarithm of a one-dimensional density function. Given an integer $k > 2$, the lower bound of data L, the upper bound of data U (L, U can be infinite), and a sequence $t_1, ..., t_k$ with $L < t_1 < .. .< t_k < U$, we consider the space S consisting of the twice-continuously differentiable function s on (L,U), such that the restrictions of s to $[t_1, t_2]$, ..., $[t_{k-1}, t_k]$ are cubic polynomials, and are linear on $(L, t_1]$ and $[t_k, U)$. Functions of S are called natural (cubic) splines. Let $1, B_1, ..., B_{k-1}$ be a set of basis functions that span the space S. Given $\underline{\theta} = [\theta_1, ..., \theta_k]^t$ a k-dimensional vector such that:

$$\int_L^U \exp\left(\theta_1 B_1(y) + ... + \theta_k B_k(y)\right) dy < \infty \quad (15)$$

we can consider the exponential family of distributions based on this basis function :

$$f(y, \underline{\theta}) = \exp\left(\begin{array}{l} \theta_1 B_1(y) + ... \\ ... + \theta_k B_k(y) - C(\underline{\theta}) \end{array} \right) \quad (16)$$

where $C(\underline{\theta})$ is a normalising constant such that:

$$\int_{\mathbb{R}} f(y, \underline{\theta}) dy = 1 \quad (17)$$

We note Θ the space of all θ which verify the above constraints. Let consider N sample data $y_1, ..., y_N$ from the distribution we want estimate the density. The log-likelihood function corresponding

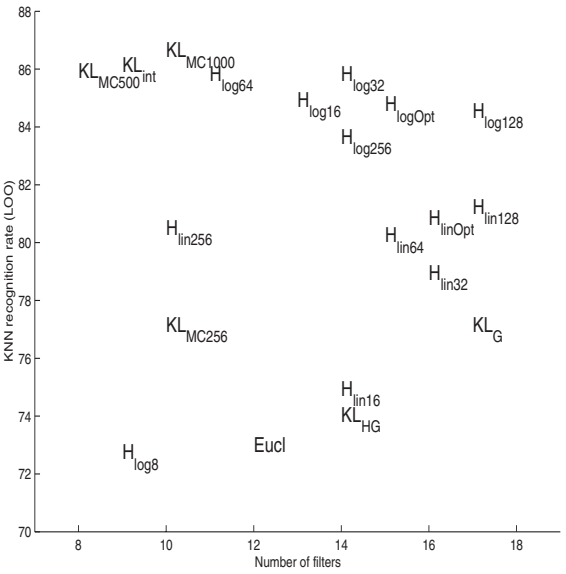


Fig. 4. Results of the "Leave One Out" classification according to the signature/distance used.

to the logspline family is:

$$L(\underline{\theta}) = \sum_{i=1}^{N} \log\left(f(y_i, \underline{\theta})\right), \qquad \underline{\theta} \in \Theta \quad (18)$$

This function is strictly concave on Θ, so if the maximum-likelihood estimate $\hat{\underline{\theta}}$ exists, it is unique. When $(L,t_1]$ and $[t_k,U)$ contain at least one sample, and the other intervals contain at least four values, the maximum-likelihood estimate $\hat{\underline{\theta}}$ exists, is unique, and we refer to :

$$\hat{f}(.) = f(.;\hat{\underline{\theta}}) \quad (19)$$

as the logspline density estimate. Kooperberg and Stone have proposed a method and a computer code in the Splus environment to automatically determine the optimal number k, the values of the t_i, and calculate the maximum-likelihood estimate. The number of knot is chosen according to the Akaike Information Criterion (AIC), then they are placed at or near selected order statistics, *i.e.* it depends only on the ordering of the data and not on its numerical values.

3.4.2. Implementation

We use the adaptation of this code to the R environment (Ripley, Kooperberg, 2000), which estimates densities according to the method presented above. It provides densities, probabilities, quantiles and random samples from the estimated logspline densities. Two methods are implemented for the estimation of the Kullback-Leibler divergence and are described below. We refer in appendix A in an other possible implementation of the distances which requires to have access to the gradient of $\hat{\underline{\theta}}$. In the existing program, Kooperberg use another basis and various transformations for this calculus since it is not intended for direct use. A total change of the program would be required to access to this gradient, and this is not the topic of this article. The numerical processes that we propose hereinafter give an estimation of Kullback-Leibler divergence in a general case, even if one uses an other method of density estimation.

Let f_1 and f_2 two densities estimated with the logspline model for instance. We can directly use equation (3) to calculate the KL divergence. We note $KL_{int}(f_1,f_2)$ the estimate of the KL divergence with the integral formula.

Nevertheless, we can remark that equation (3) is equivalent to:

$$KL(f_1, f_2) = E_{f_1}\left[\log\left(\frac{f_1(X)}{f_2(X)}\right)\right] \quad (20)$$

where E[.] is the expectation, and **X** is a random vector which follows the law f_1. This Monte Carlo implementation can be calculated by the natural estimate of the expectation (law of large numbers) :

$$KL_{MC}(f_1, f_2) = \sum_{k=1}^{p} \log\left(\frac{f_1(x_k)}{f_2(x_k)}\right) \quad (21)$$

where the x_k are p random sample from the density f_1. We directly generate these samples and the value of the densities at these points from the program of Kooperberg (Ripley & Kooperberg, 2000).

4. Empirical results

4.1 Classification paradigm

In the following we compare the efficiency of the pairs signatures/distances, using a simple K Nearest Neighbours (KNN) classifier. We calculate the responses of the 540 images to the twenty most dispersed filters (Le Borgne & Guérin Dugué, 2001), and we compute all the signatures we have described above: mean value, histograms, logspline densities... Several values were tested for the parameter K, ranging from 1 to 19, and the best was retained.

The efficiency of each strategy is evaluated by the average of the trace of the confusion matrix. In a classification paradigm, construction of this matrix is a critical point, because the "true confusion matrix" is always unknown, and we can only calculate an "apparent confusion matrix" using repeated train-and-test partitions of data. Several resampling methods exist and provide more or less biased and variable estimates of the true recognition rate. The choice of the method mainly depends on the number of available sample, and the accuracy we want. In this paper, unless our database is not so small, we use two computationally expensive methods which are "leave-one-out" and "bootstrap", in order to limit bias and variance of our results.

The "leave-one-out" resampling method for N images consists on N train-and-test classifications with N-1 images for learning and 1 image for testing. Thus, it is a cross-validation method which produces too wide confidence intervals for the true error rate. Since the size of our database is not very large (540 images), we also validate our results with a bootstrap

procedure which gives much narrower confidence limits (Henery, 1994).

The "bootstrap resampling method" (Efron & Tibshirani, 1993) for N images consists on classifying N_B sets of images of size N_L for learning, and N_T=N-N_L sets for testing. The bootstrap estimator of the true recognition rate is the average of the N_B recognition rate. The variance of these N_B classifications gives an indication about the variability of the result. In our case we have chosen $N_L = N_T = 540 / 2 = 270$, because it realises the best compromise between bias and variance (Burman, 1989).

4.2 Results

Results of KNN classifications for the different models are reported in figure 4. Evaluation of performance has been estimated with a "leave one out" process, with an optimal value k among {1,3,...,19}. For all the signatures and their corresponding distances it indicates the best recognition rate we obtain, and the corresponding number of ICA filters. One and two parameters models give the results respectively indicated by "Eucl", "KL_{HG}" (for half-Gaussian fit) and "KL_G" (for Kullback Leibler of Gaussian pdf). "H_{linN}" is a histogram with N bins of equal width, and "H_{logN}" is the histogram for N bins in a logarithmic scale. The estimation of the number of bin with equation (12) is 65 at the minimum, 82 at the maximum, and 75 in average, according to the filter we consider. We have reported the results of classification with an optimisation of the number of bins (in the sense of equation (12)) as "H_{linOpt}" and "H_{logOpt}" in figure 4. "KL_{int}" indicates a Kullback Leibler divergence between two logspline-modelled densities implemented with the integral formula, and KL_{MCp} is the same with the p sample Monte Carlo implementation. Note we always use a symmetric version of the Kullback Leibler divergence.

The recognition rate increases with the accuracy of the model of signature. Simple mean leads to the weakest recognition rate (less than 74%), half-Gaussian fit is about 74%, while it increases at 78% for the two-parameters model (mean and variance), about 80% for histograms with linear distribution of bins and more than 85% for the logspline estimation of density. As it is illustrated in figure 5, when model become more complex, we better fit the queue of the distributions. This result suggests that the most informative part of the distributions for discriminating images with ICA filters is localised in the queues.

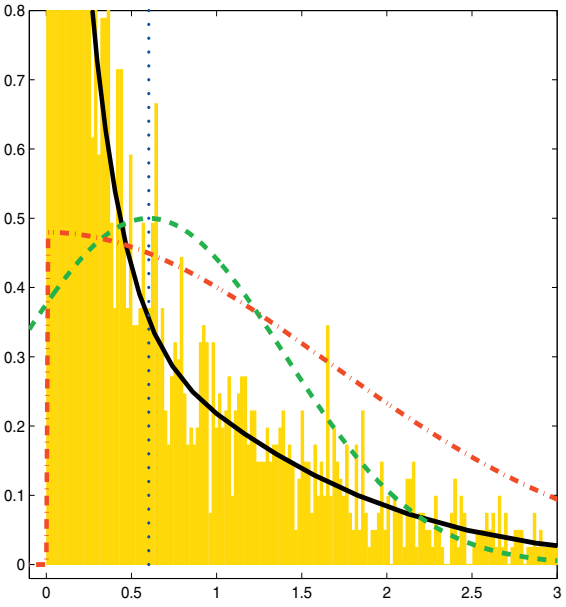


Fig. 5. Fit of data with several models - Dotted: average - Dash-Dot: half-normal - Dashed line: gaussian (2 parameters) - Solid line: logspline.

Integral implementation ("KL_{int}") gives about same results as the Monte Carlo one. For this later, 500 samples are enough to obtain a recognition rate of more than 85%. Concerning histograms, with a linear distribution of bins, the number of bins could vary from 256 to 64 (and even 32) without any significant loss of performance. We remark that optimisation of the number of bin for each filter (response H_{linOpt}) give a similar result. As expected, a signature with logarithmic distribution of bins increases meaningfully the classification results, since it is more adjusted to the sparse responses of ICA filters. It also allows the use of less complex histograms with only 32 or 16 bins without significant loss of performance since it takes advantage of the *a priori* knowledge about the sparseness of the densities. This signature is interesting from a practical point of view, leading to recognition rates almost as good as those obtained with the logspline model.

We can remark the little number of descriptor we use to reach all these performances (9 to 17). The criterion that determines the choice of filters is discussed in (Le Borgne & Guérin-Dugué, 2001).

For the best results, we also classify using the bootstrap process, in order to become independent from the learning database. We use 100 bootstrap samples, and we report the average and standard deviation of the resulting classifications in table 1. Hence, logspline density estimation reaches the best recognition rate with more than 82.5%. We remark that for monte carlo implementation, 500 samples are sufficient to reach the best recognition rates.

All these experiment were conduced in a Matlab environment. Computationally speaking, one or two parameters models are largely less greedy than others, since we rapidly can compute signature, but above all, distances can be computed all together with a simple inner product. For other models, we have to compute the distance for each couple of image.

Distance	100 bootstrap samples	
	μ_{Boot} (%)	σ_{Boot} (%)
KL_{MC_1000}	82,5	1,6
KL_{MC_500}	82,8	1,8
KL_{int}	82,6	1.8
H_{Log_32}	81,8	1,8

Table 1. Average (μ_{Boot}) and standard deviation (σ_{Boot})of the KNN recognition rate after a bootstrap resampling with 100 samples, for the four best results of the LOO classification (see text for details).

5. Conclusion

In this paper, we have presented several models for the responses of images to ICA filters, and the way to compute the Kullback-Leibler divergence with these models, as a measure of similarity between images. As a consequence of the chosen method, Independent Component Analysis, we can fully take advantage of the Kullback-Leibler divergence, computing as the sum of the KL divergence between the marginal densities.

The evaluation of performance was done with a KNN classification paradigm, validated by a leave-one-out and a bootstrap resampling. The results show that the recognition rates increase with the capacity of the models to well fit the queues of the distributions, but the computation of the signatures and distances between images is also more and more complex and computational demanding. Nevertheless, the method we propose in the appendix A could significantly reduce the computing cost for distances. If we take advantage of the *a priori* knowledge we have about the responses of ICA filters to natural images, we estimate the density of the logarithm of data, that leads to recognition rates almost as good as the logspline based signature.

This paper focuses on feature extraction with ICA, using these for defining similarity between images. It willingly ignores the scale ability issues that one can meet in the design of a real CBIR system, since it concerns our future works. Even if the bootstrap resampling allows claiming a kind of independence with respect to the choice of the learning and the testing database, we expect that performance will decrease when we will index several thousands of images. We will confront with two different kinds of problem. The first deals with combinational complexity when we increase the number of images without make the class more complex. The risk in that case is that the K-nearest-neighbour classifier fails, and a solution would be to opt for a classifier that defines prototypes of class. The second problem deals with the increasing of intrinsic complexity of classes. Our framework could then have to be extended and include other attributes, such as colour (Vailaya *et al*., 2001) or statistical context (Torralba & Sinha, 2001). Since the task of discrimination will claim more precision in designing the features, we expect that the logspline model will lead to the best performances.

Acknowledgements

The authors wish thank professor Erkki Oja and Jorma Laaksonen for the welcome in the Laboratory of Computer and Information Science of Helsinky. We also thank Aapo Hyvärinen, Patrick Hoyer, Jarmo Hurri and Mika Inki for fruitful discussions about ICA and extraction of features. Finally, we thank the two anonymous referees who pointed out several weakness in the first version of the paper, and thus helped improve the present manuscript. The Rhône-Alpes region funds Hervé Le Borgne in the "ASCII" project on image indexing. A part of this work was fund by the Elesa-Imag project "SASI" on advanced statistics for signals and images, and "SCOPIE" on perception and image retrieval systems.

6. Appendix A : Kullback-Leibler divergence and logspline model

Logspline model provides an elegant formula to estimate Kullback-Leibler divergence (relative entropy) between two density functions. Let f_1 and f_2 be two densities estimated on the same spline basis functions B(x) According to paragraph 3.4.1, we write the densities as:

$$f_i(y,\underline{\theta}) = \exp\left(\left\langle \underline{\theta}_i, B(y)\right\rangle - C(\underline{\theta}_i)\right) \qquad i \in \{1,2\}$$

with the logspline coefficients and basis :

$$B(y)=\left(1,\mathrm{B}_1(\mathrm{y}), ..., \mathrm{B}_{\mathrm{k\text{-}1}}(\mathrm{y})\right)$$

$$\underline{\theta}_i = \left(\theta_{i,1},\theta_{i,2},...,\theta_{i,k-1}\right)$$

Then the kullback Leibler divergence between f_1 and f_2 is :

$$KL(f_1,f_2) = \int_{\mathbb{R}} f_1 \log\left(\frac{f_1}{f_2}\right) = E_{f_1}\left[\log\left(\frac{f_1}{f_2}\right)\right]$$

Since the densities are under an exponential form:

$$KL(f_1,f_2) = E_{f_1}\left[\left\langle \underline{\theta}_1 - \underline{\theta}_2, B(y)\right\rangle - C(\underline{\theta}_1) + C(\underline{\theta}_2)\right]$$

$$KL(f_1,f_2) = \left\langle \underline{\theta}_1 - \underline{\theta}_2, E_{f_1}\left[B(y)\right]\right\rangle - C(\underline{\theta}_1) + C(\underline{\theta}_2)$$

Remember that at convergence (at $\theta = \hat{\theta}$), the loglikelihood is maximum, so its derivate is zero. Let Y_1,... Y_n be a random sample of size n from f_1 :

$$\frac{\partial L}{\partial \theta_j}(\hat{\theta}) = 0 = \sum_{i=1}^{n} B_j(Y_i) - n\frac{\partial C}{\partial \theta_j}(\hat{\theta})$$

So :

$$E_{f_1}\left[B(y)\right] = grad\left[C(\underline{\theta})\right]_{\underline{\theta}=\underline{\theta}_1}$$

And finally we have:

$$KL(f_1,f_2) = \left\langle \underline{\theta}_1 - \underline{\theta}_2, grad\left[C(\underline{\theta})\right]_{\underline{\theta}=\underline{\theta}_1}\right\rangle - C(\underline{\theta}_1) + C(\underline{\theta}_2)$$

References

Amari S., Cichocki A., Yang H.H., 1996. A new learning algorithm for blind signal separation. In D.S. Touretsky, M.C. Mozer, & M.E. Hasselmo, Advances in neural information processing systems, 8, pp 757-763. Cambridge, MA : MIT press.

Basseville M., 1996. Information: entropies, divergences et moyennes (In French). Research Report IRISA no 1020.

Bell A.J., Sejnowsky T. J., 1997. The Independent Components of Natural Scenes are Edge Filter, Vision Research, vol. 36, pp. 287-314.

Burman, P., 1989. A comparative study of ordinary cross-validation, v-fold cross validation and the repeated learning testing methods, Biometrika, 76 (3), 503-514.

Comon, P., 1994. Independant Component Analysis – a new concept ?, Signal processing, vol 36, pp 287-314

Cox,I.J., Miller M.L., Omohundro, Yianilos P.L.. "PicHunter: Bayesian Relevance Feedback for Image Retrieval", Int. Conf. On Pattern Recognition, Austria, 1996.

Del Bimbo, A., 1999. Visual Information Retrieval, M. Kaufmann Ed, San Francisco, USA..

Do, M.N., Vetterli, M., 2002. Wavelet-Based Texture Retrieval Using Generalised Gaussian Density and Kullback-Leibler Distance, IEEE trans. on image processing, vol 11, N° 2, pp. 146-158.

Efron B., Tibschirani R.J., 1993. An introduction to the bootstrap. Monographs on statistics and Applied Probability. Chapman & Hall, New-York.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press

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

Guyader, N., Le Borgne, H., Hérault J., Guérin-Dugué A., 2002. Toward the introduction of human perception in a natural scene classification system. In : IEEE workshop on Neural Networks for Signal Processing XII, pp. 385-394, Martigny, Switzerland.

Henery, R. J., 1994. Methods for Comparison. In : Michie, D., Spiegelhalter, D. J., and Taylor, C. C., editors: Machine learning, neural and statistical classification. Ellis Horwood.

Hérault J., Jutten C., Ans B., 1985, Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage nono supervisé. Proc. GRETSI, pp 1017-1020, Nice, France.

Hérault J., 2001. De la rétine biologique aux circuits neuromorphiques, chap. 3, in "Les systèmes de vision", J.M. Jolion ed., IC2 col., Hermes, Paris.

Hurri J., 1997. Independent component analysis of image data. Master's thesis, Helsinky University of Technology, Espoo, Finland.

Hyvärinen A., Oja.E., 1997. A Fast fixed-point algorithm for Independent Component Analysis, Neural Computation, vol 9, no 7, pp. 1483-1492.

Hyvärinen A., Karhunen, J., Oja, E., 2001a. Independent Component Analysis, John Wiley & Sons.

Hyvärinen A., Hoyer P., Oja E., 2001b, Image Denoising by Sparse Code Shrinkage. In S. Haykin and B. Kosko (eds), *Intelligent Signal Processing*, IEEE Press.

Izenman, A.J., 1991. Recent developments in non parametric density estimation. Journal of the American Statistical Association, 86 (413), pp 205-224.

Johansson B., 2002. A Survey on: Contents Based Search in Image Databases. http://www.isy.liu.se/cvl/Projects/VISIT-bjojo/survey/surveyonCBIR/index.html

Kooperberg, C., Stone, C.J., 1992. Logspline density estimation for censored data. Journal of Computational and Graphical Statistics, 1, 301-328.

Labbi A., Bosch H., Pellegrini, Ch. (1999). Image Categorization using Independant Component Analysis. ACAI Workshop on Biologically Inspired Machine Learning, BIML'99, July 14 (invited talk), Crete, Greece

Le Borgne H., Guérin-Dugué A., 2001. Sparse-Dispersed Coding and Images Discrimination with Independent Component Analysis. In: third International Conference on ICA and BSS, San Diego, California, USA, December 9-12, 2001.

Le Borgne H., Guyader N., Guérin-Dugué A., Hérault J., 2003. Proceedings of the seventh International Symposium on Signal Processing and its Applications ISSPA'03, vol 2, pp. 251-254, Paris, France.

Oliva, A., Torralba, A., Guerin-Dugue, A & Herault, J. 1999. Global semantic classification of scenes using power spectrum templates. Proceedings of The Challenge of Image Retrieval (CIR99), Springer Verlag BCS Electronic Workshops in Computing series, Newcastle, UK.

Olshausen B. A., Field D. J., 1997. Sparse Coding with an Overcomplete Basis Set: A strategy Employed by V1 ?, Vision Research, vol. 37, nº23, pp. 3311-3325

Pham D.T., Garrat P., Jutten C., 1992. Separation of a mixture of independent sources through a maximum likelihood approach, Proc. EUSIPCO, pp 771-774.

Puzicha, J., Rubner, Y., Tomasi, C., Buhmann, J.M., 1999. Empirical evaluation of dissimilarity measures for color and texture. International Conference on Computer Vision, pages 1165--1173. Kerkyra (Corfu), Greece.

Ripley B., Kooperberg C.L., december 19, 2000. The Logspline Package found on the CRAN project, http://lib.stat.cmu.edu/R/CRAN/

Rogowitz, B. Frese, T., Smith J., Bouman, C.A., Kalin, E., 1998. Perceptual image similarity experiments, Human Vision and Electronic Imaging III, Proc. of the SPIE, vol 3299, pp. 576-590, San Jose, CA.

Silverman B.W., 1986. Density estimation for statistics and data anlysis, Chapman and Hall, London.

Stricker, M., and Orengo, M., 1995. Similarity of Color Image",in Storage and Retrieval for Image and Video Databases, Proc. SPIE 2420, pp 381-392, 1995.

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

Torralba, A., Sinha, P., 2001. Statistical context priming for object detection. CBCL Paper #205/AI Memo #2001-020, Massachusetts Institute of Technology, Cambridge, MA, September 2001

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

Vailaya A., Figueiredo M., Jain A., Zhang H.J., 2001. Image classification for Content-Based Indexing, IEEE transaction on Image Processing, vol 10, n° 1, pp. 117-130.

Van der Schaaf A., Van Hateren J. H., 1996. Modelling the power spectra of natural images: statistics and information. Vision Research, vol. 36, pp. 2759-2770.

Van Hateren J.H., Van der Schaaf A., 1998. Independent component filters of natural images compared with simple cells in primary visual cortex, Proc. of the Royal Soc. of London, series B, vol 265, pp. 359-366

Willmore B., Watters P. A., Tolhurst D. V. 2000. A comparison of natural-image-based models of simple-cell coding, Perception, vol. 29, pp. 1017-1040