Preprint - Published in Springer Verlag LNCS 5416, p. 284-304, 2009

Sparse Multiscale Patches for Image Processing

Paolo Piro, Sandrine Anthoine, Eric Debreuve, and Michel Barlaud

Université de Nice Sophia-Antipolis / CNRS, Sophia-Antipolis, France

Abstract. This paper presents a framework to define an objective measure of the similarity (or dissimilarity) between two images for image processing. The problem is twofold: 1) define a set of features that capture the information contained in the image relevant for the given task and 2) define a similarity measure in this feature space.
In this paper, we propose a feature space as well as a statistical measure on this space. Our feature space is based on a global descriptor of the image in a multiscale transformed domain. After decomposition into a Laplacian pyramid, the coefficients are arranged in intrascale/interscale/interchannel patches which reflect the dependencies between neighboring coefficients in presence of specific structures or textures. At each scale, the probability density function (pdf) of these patches is used as a descriptor of the relevant information. Because of the sparsity of the multiscale transform, the most significant patches, called *Sparse Multiscale Patches (SMP)*, characterize efficiently these pdfs. We propose a statistical measure (the Kullback-Leibler divergence) based on the comparison of these probability density functions. Interestingly, this measure is estimated via the nonparametric, k-th nearest neighbor framework without explicitly building the pdfs.
This framework is applied to a query-by-example image retrieval task. Experiments on two publicly available databases showed the potential of our *SMP* approach. In particular, it performed comparably to a *SIFT*-based retrieval method and two versions of a fuzzy segmentation-based method (the *UFM* and *CLUE* methods), and it exhibited some robustness to different geometric and radiometric deformations of the images.

Key words: multiscale transform, sparsity, patches, Kullback-Leibler divergence, k-th nearest neighbor.

1 Introduction

1.1 Similarity in image processing

Defining an objective measure of the similarity (or dissimilarity) between two images (or parts of them) is a recurrent question in image processing that is dealt with in quite different ways. When dealing with inverse problems such as denoising or deconvolution of images, a similarity measure is needed to evaluate how well the estimate explains the observations. However, for these problems,

efforts have been concentrated in the conditioning of the inverse operator as well as the spatial properties of the estimated images. The measure of fitness to the data has been less studied and is usually a simple euclidean norm in pixel space such as: $d(I_1, I_2) = \sqrt{\sum_{i \in \{pixel\}} |I_1(i) - I_2(i)|^2}$. At the other end of the spectrum, for some applications, the similarity measure is at the core of the problem and has received much more attention. This is the case for applications such as tracking or image retrieval, where the task is to rank the images of a database according to their visual similarity to the given query image. In any case, defining a similarity measure is a two steps process:

1. Define a set of properties that capture the information contained in the image relevant for the given task. This step defines the so-called feature space.
2. Define a similarity measure in the feature space. This measure is often (but not always) a distance.

Number of possibilities have been explored for the feature space itself. Some spaces involve a transform domain (e.g. wavelet transforms), some are based on various descriptors. A variety of descriptors (see [1] for a review) has been proposed in the literature. Local descriptors (e.g. salient points [2]) are based on a subset of the pixels of the image while global descriptors give information about the image as a whole (e.g. color histograms [3]). Local descriptors exploit the information given by a limited of number of points of interest together with their spatial neighborhood. Hence much information in the image is not used in these methods (see [4] for an extensive comparison and performance evaluation of most local descriptors). On the contrary, global descriptors include information of the whole image (e.g. histograms of intensity values). Global descriptors may be defined at the pixel level (e.g. color histograms [3]) and include no notion of spatial correlation or at the patch level including spatial and/or scale correlations. The concept of global patch descriptors is supported by statistical studies on images [5]. Here, we propose a new descriptor of this kind.

The similarity measure can range from simple euclidean norm to more sophisticated measures: robust estimators have been used for optical flow [6, 7], Bhattacharya's distance for tracking [8], entropic measure such as entropy, Kullback-Leibler divergence or mutual information for registration [9, 10]. A general requirement for the similarity measure is the visual relevance, i.e. a strong correlation with human perception of similarity itself. Research in vision science has already brought some perspectives on how to do so [11]. Nevertheless, designing systems purely based on the perceptual characteristics of the human visual system is a difficult task. Therefore, once meaningful features have been selected, we prefer to employ metrics that have a mathematical foundation and can be easily implemented. For this purpose, several distance metrics have been used to compare feature vectors for various tasks of image processing. The authors of [12] give a variety of such measures and empirically show how the selection of a metric affects the performances of a retrieval system.

In this paper we propose a feature space as well as a statistical measure on this space. Our feature space is based on a global descriptor in a transformed

domain that we call *Sparse Multiscale Patches*. The measure we propose on this space is statistical: it compares the probability density function (pdf) of these patches.

1.2 Proposed feature space and measure

We propose a new descriptor based on *Sparse Multiscale Patches*. In short, we integrate using probability distributions the local information brought by the *SMP*. The key aspects of these descriptors are the following:

- A *multiscale* representation of the images;
- Inter/intrascale patches that describe locally the structure of the image at a given scale;
- A *sparse* repartition: most of the energy is concentrated in a few patches.

Note that the occurrence in different parts of an image of similar patches of spatially neighboring pixels has been exploited in image processing [13–15]. Here the concept is used for multiscale coefficients as proposed in [16].

The visual content of images is represented by patches of multiresolution coefficients. The extracted feature vectors are viewed as samples from an unknown multidimensional distribution. The multiscale transform of an image being sparse, a reduced number of patches yields a good characterization of the distribution. We estimate the similarity between images by a pseudo-distance (or measure) between these multidimensional probability density functions.

We propose to use the Kullback-Leibler (KL) divergence as a similarity measure that quantifies the closeness between two probability density functions. Such measure has already shown good performances in the context of image retrieval [12]. It has already been used for the simple case of parametrized marginal distributions of wavelet coefficients [17, 18], assuming independence of the coefficients. In contrast, we define multidimensional feature vectors (patches), that capture interscale and intrascale dependencies among subband coefficients. These are better adapted to the description of local image structures and texture. In addition, for color images, we take into account the dependencies among the three color channels; hence patches of coefficients are also interchannel. This approach implies to estimate distributions in a high-dimensional statistical space, where fixed size kernel options to estimate distributions or divergences fail. Alternatively, we propose to estimate the KL divergence directly from the samples by using the k-th nearest neighbor (kNN) approach, *i.e.* adapting to the local sample density.

1.3 Organization of the paper

This paper is organized as follows. In Section 2, we define our feature space which consists of inter/intrascale and interchannel patches of Laplacian pyramid coefficients for color images, called *Sparse Multiscale Patches*. We then define the global similarity on this feature space in Section 3 by combining similarities

between the probability density functions of these patches at different scales. The comparison between pdfs is measured by the KL divergence. We also explain how to estimate this quantity. Finally, in the last section we illustrate the use of the proposed measure in a particular application: image retrieval.

2 Feature space: Sparse Multiscale Patches

Throughout this paper, we will denote the input image by I, the scale of the multiresolution decomposition by j, and the location in the 2D image space by k.

2.1 Multiscale coefficients: strengths and limits

The wavelet transform enjoys several properties that have made it quite successful in signal processing and that are relevant for the definition of similarity between images. Indeed, it provides a sparse representation of images, meaning that it concentrates the informational content of an image into few coefficients of large amplitude while the rest of the coefficients are small. This combined with a fast transform is what makes wavelet thresholding methods so powerful: in fact just identifying large coefficients is sufficient to extract where the information lies in the image. For example, denoising can be done very efficiently by simply thresholding wavelet coefficients as proved in [19]. Such simple coefficient-wise treatments provide results of excellent quality at a reduced computational cost.

In fact, these classical wavelet methods treat each coefficient separately, relying on the fact that they are decorrelated. However, the wavelet coefficients are not independent and these dependencies are the signature of structures present in the image. For example, a discontinuity between smooth regions at point k_0 will give large coefficients at this point at all scales j ($w(\mathrm{I})_{j,k_0}$ large for all j). Classical methods using coefficient-wise treatments may destroy these dependencies between coefficients and hence alter the local structure of images. Therefore models using the dependencies between coefficients have been proposed and used in image enhancement (e.g. [16, 20]). In particular, the authors of [16] introduced the concept of patches of wavelet coefficients (which they called "neighborhoods of wavelet coefficient") to represent efficiently fine spatial structures in images.

2.2 Multiscale patches for color images

Following these ideas, we define a feature space based on a sparse descriptor of the image content by a multiresolution decomposition. More precisely, we group the Laplacian pyramid coefficients of the three color channels of the image I into coherent sets called patches. Here the coherence is sought by grouping coefficients linked to a particular scale j and location k in the image.

In fact, the most significant dependencies are seen between a coefficient $w(\mathrm{I})_{j,k}$ and its closest neighbors in space: $w(\mathrm{I})_{j,k\pm(0,1)}$, $w(\mathrm{I})_{j,k\pm(1,0)}$ and in scale:

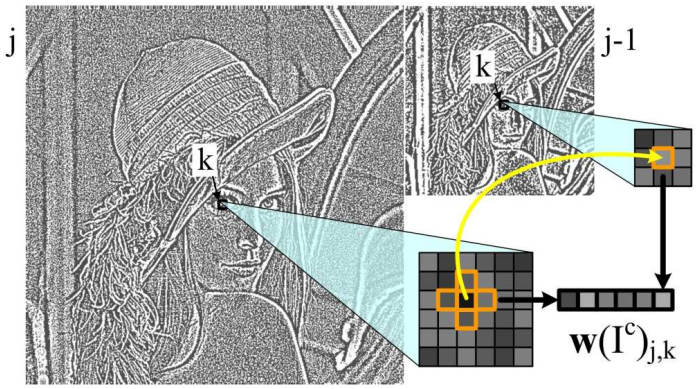

Fig. 1. Building a patch of multiscale coefficients, for a single color channel image.

$w(\mathrm{I})_{j-1,k}$, where $j-1$ represents the scale a step coarser than the scale j. Grouping the closest neighbors in scale and space of the coefficient $w(\mathrm{I})_{j,k}$ in a vector, we obtain the patch $\vec{w}(\mathrm{I})_{j,k}$ (see Fig. 1):

$$\vec{w}(\mathrm{I})_{j,k} = \left(w(\mathrm{I})_{j,k}, w(\mathrm{I})_{j,k\pm(1,0)}, w(\mathrm{I})_{j,k\pm(0,1)}, w(\mathrm{I})_{j-1,k}\right) \tag{1}$$

which describes the structure of the grayscale image I at scale j and location k. The probability density functions of such patches at each scale j have proved to characterize fine spatial structures in grayscale images [16, 21]. Such patches are therefore relevant features for our problem as will be seen in Section 4.3.

We consider color images in the luminance/chrominance space: $\mathrm{I} = (\mathrm{I}^Y, \mathrm{I}^U, \mathrm{I}^V)$. Since the coefficients are correlated through channels, we aggregate in the patch the coefficients of the three channels:

$$\mathbf{w}(\mathrm{I})_{j,k} = \left(\vec{w}(\mathrm{I}^Y)_{j,k}, \vec{w}(\mathrm{I}^U)_{j,k}, \vec{w}(\mathrm{I}^V)_{j,k}\right) \tag{2}$$

with $\vec{w}(\mathrm{I}^Y)_{j,k}$, $\vec{w}(\mathrm{I}^U)_{j,k}$ and $\vec{w}(\mathrm{I}^V)_{j,k}$ given by Eq.(1).

The low-frequency approximation that results from the Laplacian pyramid is also used to build additional feature vectors. Namely, 3×3 pixel neighborhoods along all three channels are joined together to form patches of dimension 27 (whereas patches from the higher-frequency subbands are of dimension 18, as defined in Eq.(2)). The union of the higher-frequency and low-frequency patches forms our feature space. The patches of this augmented feature space will be denoted by $\mathbf{w}(\mathrm{I})_{j,k}$.

2.3 Multiscale transform

The coefficients are obtained by a Laplacian pyramid decomposition [22]. Indeed, critically sampled tensor wavelet transforms lack rotation and translation

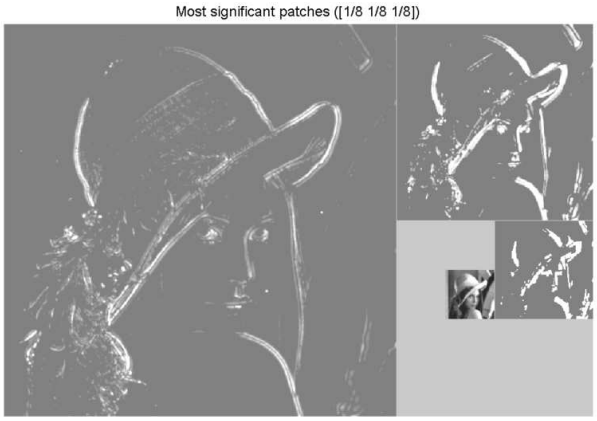

Fig. 2. White indicates the location of patches of largest energy (1/8 of the patches is selected for each subband).

invariance and so would the patches made of such coefficients. Hence we prefer to use the Laplacian pyramid which shares the sparsity and inter/intrascale dependency properties with the wavelet transform while being more robust to rotations. Moreover, the Laplacian pyramid is at the basis of the SVC standard and thus our approach will be compatible with SVC.

2.4 Sparsity of the multiscale patches

As we have seen earlier, multiscale coefficients provide a sparse representation of images by concentrating the information into a few coefficients of large amplitude and this sparsity is exploited on the raw coefficients in thresholding methods. As illustrated in Fig. 2, our experiments show that the patches of multiscale coefficients of large overall energy (sum of the square of all coefficients in a patch) also concentrate the information. Since the total number of patches in an image decomposition is $4/3N$ with N the number of pixels in the image, the number of samples we have in the feature space is quite large as far as measuring a similarity is concerned. The possibility of selecting a small number of patches which represent the whole set well is therefore highly desirable. In practice, we selected a fixed proportion of patches at each scale of the decomposition and proved that the resulting similarity measure (defined in section 3) remains consistent (see [23] for details). This is exploited to speed up our computations.

Note that other selecting procedures may be investigated such as using the energy of the central coefficient, using the sum of absolute differences in the patches or thresholding based on the variance of the patches.

Let us now explain how we define a similarity in this feature space.

3 Similarity measure

3.1 Definition

Our goal is to define a similarity measure between two images I_1 and I_2 from their feature set i.e. from their respective set of patches $\{\mathbf{w}(\mathrm{I}_1)_{j,k}\}_{j,k}$ and $\{\mathbf{w}(\mathrm{I}_2)_{j,k}\}_{j,k}$. When images are clearly similar (e.g. different views of the same scene, images containing similar objects...), their patches $\mathbf{w}(\mathrm{I}_1)_{j_l,k_l}$ and $\mathbf{w}(\mathrm{I}_2)_{j_l,k_l}$ do not necessarily correspond. Hence a measure comparing geometrically corresponding patches would not be robust to geometric transformations. Thus, we propose to compare the pdfs of these patches. Specifically, for an image I, we consider for each scale j the pdf $p_j(\mathrm{I})$ of the set of patches $\{\mathbf{w}(\mathrm{I})_{j,k}\}_k$.

To compare two pdfs, we place ourselves in the framework of Bregman divergences, which allows to generate a class of pseudo-metrics that generalize the classical squared Euclidean distance. These divergences do not necessarily satisfy the triangle inequality nor are symmetric (they are not metrics) but share similar properties. A Bregman divergence is derived from a convex function. For example, the square Euclidean distance stems from the square function $f(x) = x^2$, while the Kullback-Leibler divergence derives from the function $f(x) = x \log x$ [24, 25]. In this paper, we use the Kullback-Leibler divergence because it is a Bregman divergence that derives from the Shannon differential entropy (quantifies the amount of information in a random variable through its pdf). The Kullback-Leibler divergence (D_{kl}) is the following quantity [12]:

$$D_{kl}(p_1||p_2) = \int p_1 \log(p_1/p_2), \tag{3}$$

This divergence has been successfully used for other applications in image processing in the pixel domain [26, 15], as well as for evaluating the similarity between images using the marginal pdf of the wavelet coefficients [17, 18]. In this paper, we propose to measure the similarity $S(\mathrm{I}_1, \mathrm{I}_2)$ between two images I_1 and I_2 by summing over scales the divergences between the pdfs $p_j(\mathrm{I}_1)$ and $p_j(\mathrm{I}_2)$:

$$S(\mathrm{I}_1, \mathrm{I}_2) = \sum_j \alpha_j D_{kl}(p_j(\mathrm{I}_1)||p_j(\mathrm{I}_2)) \tag{4}$$

where α_j is a positive weight that may normalize the contribution of the different scales.

3.2 Limits of the parametric approaches to the estimation

The estimation of the similarity measure S consists of the evaluation of divergences between pdfs $p_j(\mathrm{I_i})$ of high dimension. This raises two problems. Firstly, estimating the KL divergence, even with a good estimate of the pdfs, is hard because this is an integral in high dimension involving unstable logarithm terms. Secondly, the accurate estimation of a pdf itself is difficult due to the lack of samples in high dimension (curse of dimensionality). The two problems should be embraced together to avoid cumulating both kinds of errors.

A first idea consists in parametrizing the shape of the pdf. The marginal pdf of multiscale coefficients is well modeled by generalized Gaussians. In this case, the KL divergence is an analytic function of the pdf parameters. This technique has been used in [17, 18] to compare images on the basis of the marginal pdf of their wavelet coefficients. To our knowledge, the generalized Gaussian model cannot be extended to account for correlations in higher dimension. Mixture of Gaussians on the other hand are efficient multidimensional models accounting for correlations that fit well the pdf of wavelet coefficients patches [21]. However the KL divergence is not an analytic function of the model parameters.

Thus, we propose to make no hypothesis on the pdf at hand. We therefore spare the cost of fitting model parameters but we have to estimate the divergences in this non-parametric context. Conceptually, we combine the Ahmad-Lin approximation of the entropies necessary to compute the divergences with "balloon estimates" of the pdfs using the kNN approach.

3.3 Non-parametric estimation of the similarity measure

The KL divergence can be written as the difference between a cross-entropy $\mathrm{H_x}$ and an entropy H (see Eq.(3)):

$$\mathrm{H_x}(p_1, p_2) = -\int p_1 \log p_2, \qquad \mathrm{H}(p_1) = -\int p_1 \log p_1. \tag{5}$$

Let us explain how to estimate these terms from an i.i.d sample set $\mathcal{W}^1 = \{\mathbf{w}_1^1, \mathbf{w}_2^1, .., \mathbf{w}_{N_1}^1\}$ of p_1 and an i.i.d sample set $\mathcal{W}^2 = \{\mathbf{w}_1^2, \mathbf{w}_2^2, .., \mathbf{w}_{N_2}^2\}$ of p_2. (The samples are in $\mathbb{R}^d$.)

Assuming we have estimates $\widehat{p_1}$, $\widehat{p_2}$ of the pdfs p_1, p_2, we use the Ahmad-Lin entropy estimators [27]:

$$\mathrm{H_x^{al}}(\widehat{p_1}, \widehat{p_2}) = -\tfrac{1}{N_1} \sum_{n=1}^{N_1} \log[\widehat{p_2}(\mathbf{w}_n^1)], \qquad \mathrm{H^{al}}(\widehat{p_1}) = -\tfrac{1}{N_1} \sum_{n=1}^{N_1} \log[\widehat{p_1}(\mathbf{w}_n^1)]. \tag{6}$$

General non-parametric pdf estimators from samples can be written as a sum of kernels K with (possibly varying) bandwidth h (see [28] for a review):

$$\widehat{p_1}(x) = \tfrac{1}{N_1} \sum\nolimits_{n=1}^{N_1} K_{h(\mathcal{W}^1, x)}(x - \mathbf{w}_n^1). \tag{7}$$

– Parzen estimators $h(\mathcal{W}^1, x) = h$: the bandwidth is constant. They perform very well with samples in one dimension but become unadapted in high dimension due to the sparsity of the samples: the trade-off between a bandwidth large enough to perform well in low local sample density (which may *oversmooth* the estimator) and a bandwidth small enough to preserve local statistical variability (which may result in an unstable estimator) cannot always be achieved. To cope with this problem, kernel estimators using adaptive bandwidth have been proposed;

- Sample point estimators $h(\mathcal{W}^1, x) = h_{\mathcal{W}^1}(\mathbf{w}_i^1), i \in \{1, N_1\}$: the bandwidth adapts to each sample $\mathbf{w}_i^1$ given the sample set $\mathcal{W}^1$;
- Balloon estimators $h(\mathcal{W}^1, x) = h_{\mathcal{W}^1}(x)$: the bandwidth adapts to the point of estimation x given the sample set $\mathcal{W}^1$.

We use a balloon estimator with a binary kernel and a bandwidth computed in the k-th nearest neighbor (kNN) framework [28]. This is a dual approach to the fixed size kernel methods and was firstly proposed in [29]: the bandwidth adapts to the local sample density by letting the kernel contain exactly k neighbors of x among a given sample set:

$$K_{h_{\mathcal{W}}(x)}(x - \mathbf{w}_n) = \frac{1}{v_d \ \rho_{k,\mathcal{W}}^d(x)} \ \delta\Big[||x - \mathbf{w}_n|| < \rho_{k,\mathcal{W}}(x)\Big] \tag{8}$$

with v_d the volume of the unit sphere in $\mathbb{R}^d$ and $\rho_{k,\mathcal{W}}(x)$ the distance of x to its k-th nearest neighbor in $\mathcal{W}$. Although this is a biased pdf estimator (it does not integrate to one), it has proved to be efficient for high-dimensional data [28]. Plugging Eq.(8) in Eq.(6), we obtain the following estimators of the (cross-)entropy:

$$\mathrm{H}^{\mathrm{knn}}(\widehat{p_1}) = \log[N_1 \ v_d] - \log(k) + \frac{d}{N_1} \sum_{n=1}^{N_1} \big(\log[\rho_{k,\mathcal{W}^1}(\mathbf{w}_n^1)]\big), \tag{9}$$

$$\mathrm{H}_{\mathrm{x}}^{\mathrm{knn}}(\widehat{p_1}, \widehat{p_2}) = \log[N_2 \ v_d] - \log(k) + \frac{d}{N_1} \sum_{n=1}^{N_1} \big(\log[\rho_{k,\mathcal{W}^2}(\mathbf{w}_n^1)]\big). \tag{10}$$

As previously stated, these estimates are biased. A correction of the bias has been derived in [30] in a different context. In the non-biased estimators of the (cross)-entropy the digamma function $\psi(k)$ replaces the $\log(k)$ term:

$$\mathrm{H}^{\mathrm{knn}}(\widehat{p_1}) = \log[(N_1 - 1)v_d] - \psi(k) + \frac{d}{N_1} \sum_{n=1}^{N_1} \big(\log[\rho_{k,\mathcal{W}^1}(\mathbf{w}_n^1)]\big), \tag{11}$$

$$\mathrm{H}_{\mathrm{x}}^{\mathrm{knn}}(\widehat{p_1}, \widehat{p_2}) = \log[N_2 \ v_d] - \psi(k) + \frac{d}{N_1} \sum_{n=1}^{N_1} \big(\log[\rho_{k,\mathcal{W}^2}(\mathbf{w}_n^1)]\big). \tag{12}$$

And hence the KL divergence reads:

$$D_{kl}(p_1||p_2) = \log\Big[\tfrac{N_2}{N_1 - 1}\Big] + \tfrac{d}{N_1} \sum_{n=1}^{N_1} \log[\rho_{k,\mathcal{W}^2}(\mathbf{w}_n^1)] - \tfrac{d}{N_1} \sum_{n=1}^{N_1} \log[\rho_{k,\mathcal{W}^1}(\mathbf{w}_n^1)]. \tag{13}$$

This estimator is valid in any dimension d and robust to the choice of k.

4 Application: Image Retrieval

4.1 Content-based image retrieval

With the rapid growing of general-purpose image collections, performing efficiently a search on such large datasets becomes a more and more critical task.

Content-based image retrieval (CBIR) systems tackle this task by analyzing the content of images in order to provide meaningful signatures of them. Automatic search of the target images is made possible by defining a similarity measure on the underlying signature space which has a reduced dimension. As a result, content based image retrieval mainly relies on describing the image content in a relevant way (the feature space) and defining a quantitative measure on this space (the similarity measure): the retrieval task is then accomplished by ranking images in increasing order of the pseudo-distance between their feature vector and the one of a given query image.

As seen in the introduction, a variety of descriptors and similarity measures have been proposed.In this paper, we will compare our *SMP* approach to three different approaches to image retrieval, two of which share the same similarity measure. The first approach is based on *SIFT* descriptors [31], which are considered state-of-the-art amongst local descriptors. Salient points are extracted by detecting the highest coefficients in the wavelet transform of the image and *SIFT* features are then represented by histograms of the gradient orientation in regions of interest. Matching the *SIFT* features obtained in two images allows then to quantify their similarity. The other methods to which we compared ours use a segmentation-based fuzzy logic approach called *UFM* for *Unified Feature Matching* [32]. The descriptors are fuzzy features (called fuzzy sets) reflecting the color, texture, and shape of each segmented region. The *UFM* measure then integrates the fuzzy properties of all the regions to quantify the similarity between two images. Using this measure, the authors proposed two image retrieval algorithms. The first one is a strictly content-based approach (similarly to ours): it consists in ranking the database images based solely on their *UFM* distance to the query. We refer to it as the *UFM* approach. The retrieval accuracy is improved by a second method called *CLUE* : the *UFM* distances between target images themselves are used to obtain a clustering of the data from which the ranking is obtained. Consequently, this method involves additional information compared to strict content-based systems such as our approach.

4.2 Database and parameter settings

Databases

Numerical experiments were performed on two different databases. The first one contains small categories and allows to evaluate specific performances of a retrieval system such as its robustness to deformations; while the second database, with larger categories, allows to test global retrieval performances.

One of these databases contains 1,000 images of the Nister Recognition Benchmark collection [33]. The images of size 640x480 pixels are grouped by sets of four images showing the same scene or object. Their content is quite various, from indoor scenes with a single object to outdoor scenes. Images belonging to the same group are related by geometric deformations (rotation, translation, zoom and perspective) as well as radiometric deformations (changes of brightness and contrast). The ground-truth for any query image is clear: exactly the three other images of the same group are relevant.

The *SMP* retrieval method was also tested on a general-purpose image database from COREL that has been widely used for CBIR evaluation purposes. In particular, results presented in [34] can be considered as a reference. We used the same subset of the COREL database as in [34]. It includes 1,000 images of size 384×256 or 256×384 which are classified in 10 semantic categories (*Africa, Beach, Buildings, Buses, Dinosaurs, Flowers, Elephants, Horses, Food, Mountains*). In some categories, the visual similarity between two given images is not always obvious since the grouping has been made in a semantic sense (e.g. category "Africa").

Parameter settings

To build the patches as defined in section 2.2, the Laplacian pyramid was computed for each channel of the image (in the YUV color space) with a 5-point binomial filter $w_5 = [1\ 4\ 6\ 4\ 1]/16$, which is a computationally efficient approximation of the Gaussian filter classically used to build Laplacian pyramids. Three high-frequency subbands plus the low-frequency approximation were used.

In the following experiments, 1/64 (resp. 1/32, 1/16 and all) of the patches were selected in the first high-frequency (resp. second, third and low-frequency) subband to describe an image (see Section 2.4). At each scale, the KL divergence was estimated in the kNN framework, with $k = 10$. The contributions to the similarity measure from the divergences of all subbands were equally weighted ($\alpha_j = 1$ in Eq.(4)).

Note that the use of the Jensen-Shannon divergence, which is a symmetrized version of the KL divergence, has also been studied. We found that the performances of this symmetric measure are lower than with the KL divergence, and so until further understanding of this phenomenon, we report here only the results with the KL divergence.

4.3 Numerical experiments

This section presents an experimental analysis of the *SMP* method; the patch-based retrieval algorithm is evaluated in terms of its ability to retrieve similar images in a query-by-example context. Images belonging to the Nister database were used to evaluate the robustness of the method to different geometric transformations. A set of artificially-degraded images of this database was also used to evaluate the retrieval performances with respect to radiometric deformations (JPEG2000 compression noise). The global retrieval performances on the Nister database were evaluated by ROC (Receiver Operating Characteristic) curves and our method was compared to a reference *SIFT*-based retrieval algorithm. For the COREL database, the global retrieval performances were evaluated by precision curves and our method was compared with the fuzzy, segmentation-based *UFM* approach. Note that for all the following experiments, the given distance between images is S (Eq. (4)), hence the smaller the given distance is the more similar the two considered images are.

Robustness to geometric deformations

The robustness of a retrieval system to geometric deformations is its ability to find relevant images in spite of some transformations of the query, such as changes of viewpoint, rotations, zoom. This is an important requirement in image retrieval, e.g. for finding a given object in different images independently of the viewpoint. Because of its structure, the Nister database allows to evaluate the robustness of the proposed method to geometric deformations. Indeed, the database is composed of groups of four images containing the same object or scene under different viewpoints and/or lightening conditions.

Examples of retrieval for five query images taken from the database are presented in Fig. 3. In this figure, each row displays the retrieval result for the query image shown in the leftmost column. From the second column on, one can see the first 4 retrieved images ranked in increasing order of their distance to the query. Hence the second leftmost image is the most similar one, excluding the query image which is always ranked first with a distance of zero. The first retrieved images are generally relevant for the query, in spite of rotations (row 2), changes of viewpoint (rows 1, 3, 5) and zooms (rows 2, 4). This shows that the proposed descriptors and similarity measure are robust in terms of geometric deformations for the retrieval problem.

Robustness to JPEG2000 compression

Another important requirement for content-based retrieval systems is the robustness to radiometric deformations. Transmission on heterogeneous networks requires compression. This process induces a loss of quality that can be significant especially in critical transmission conditions. A retrieval system is expected to be robust to compression quality. To test the proposed method on this specific point, groups of images from the Nister database were expanded. Namely, three highly-compressed versions of one image were added to each group. They were obtained by setting three different quality levels of JPEG2000 compression.

Queries were launched on this dataset with both original and compressed images. An example of the results is shown in Fig. 4, where a non-compressed image is used as a query. The distance from the three compressed versions to the query image being quite small, the system ranked them first and before any geometrically deformed version of the query. This behavior is general and still holds when compressed images are used as queries, confirming the reliability of the proposed similarity measure in terms of its robustness to compression. Moreover, the distance to the query increases as the compression level increases. This is shown in Fig. 4, where images A, B, C are compressed versions of the query image in decreasing order of quality, the PSNR being respectively of 31.8, 29.7 and 29.3 dB.

Image retrieval performances (I): ROC curves and comparison with a *SIFT*-based method

The overall performances of the *SMP* retrieval method were evaluated by ana-

Fig. 3. Retrieval results for 5 images of the Nister database. For each row, left to right: query image; first 4 ranked images of the database (excluding the query). For each retrieved image, the distance to the query is also shown (smaller distances meaning more similar).

lyzing retrieval results on the Nister dataset; namely, each of the 1,000 images was used as a query and the similarity measure to all other images was computed. The same experiment was conducted by using a state-of-the-art retrieval method based on (local) *SIFT* descriptors [35]. For this method, the similarity measure is defined as the number of points of interest that can be matched between two images. The results of both methods were quantitatively compared by means of ROC curves. These are *recall* versus $1 - precision$ curves[1] averaged over all queries. The larger the precision and recall values, the better the retrieval performances (this corresponds to the top left side of the plot of an ROC curve).

The results of our *SMP* retrieval method are shown in Fig. 5 for different subset sizes of the database. Namely, average results on the first 100, 200 or 500 images are compared to those on the whole dataset (1000 images). Although the probability of retrieval errors increases with the size of the database, global

[1] *Recall* or *positive rate* $= \frac{D}{R}$, *1-precision* or *false positive rate* $= 1 - \frac{D}{C}$, with R=#{*relevant images for a given query*}, C=#{*desired number of retrieved images*} or *cut-off*, D=#{*correctly detected images*}.

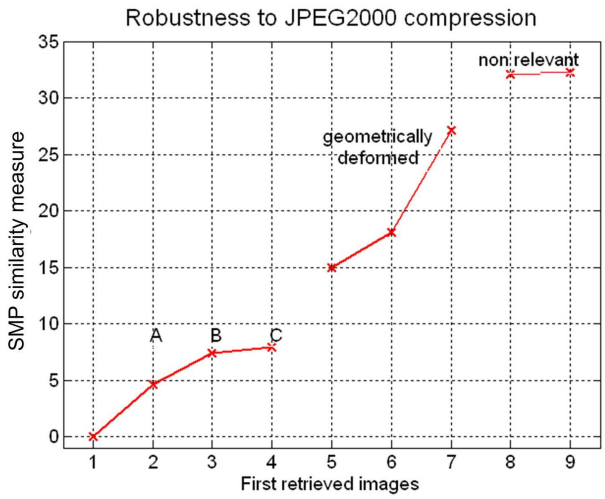

Fig. 4. Evaluation of the robustness to JPEG compression for one query image. Displayed distances are from the query to the 6 relevant images - 3 compressed (A, B, C) and 3 geometrically transformed versions of the query - and to the first 2 non-relevant images. PSNR of the compressed versions: A: 31.8dB, B: 29.7dB and C: 29.3dB.

performance is still satisfactory for a larger dataset. In any case, the best trade-off between precision and recall was reached when we retrieved three images, i.e. when the cut-off value matches exactly the number of relevant images; as a result, there is a high probability that the retrieved images are all and only the relevant ones.

Finally, the results for our *SMP* and the *SIFT*-based approach are shown in Fig. 6. The latter were obtained by running a publicly available Matlab implementation of the *SIFT* algorithm [35]. Because of the long processing time of the SIFT implementation (4.8 s on average for each comparison between two images), performing a query with each image of the database could not be done in a reasonable time. In consequence, a comparison was made by querying a subset of 100 images. In light of the ROC curves, the performances of our *SMP* method and the *SIFT*-based algorithm are comparable for this experiment.

Image retrieval performances (II): precision curve and comparison with the *UFM* method

The *SMP* retrieval method was also tested on a subset of the COREL database and compared to the *UFM* and *CLUE* methods [34]. This database is made of a small number of categories (10) containing a large number of images per category (100). Hence, ROC curves are not adapted to evaluate the global performances of a retrieval system in this case. Instead, we used the *Average*

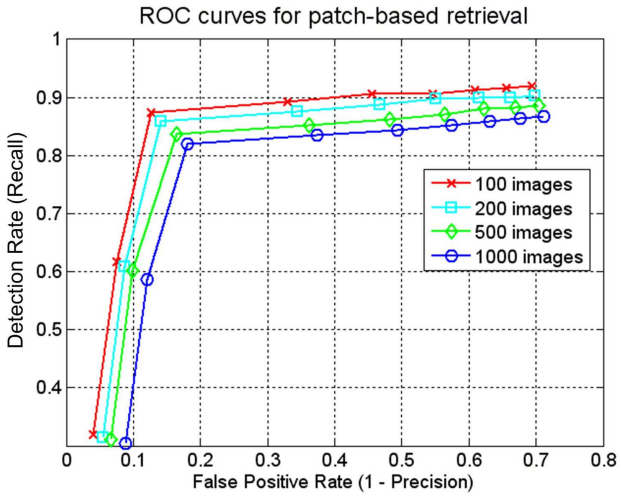

Fig. 5. Retrieval performance of the *SMP* method for different subset sizes of the Nister database; the ROC curves were obtained for cut-off values ranging from 1 to 9.

Precision to evaluate the retrieval performances for each category (the precision values for a cut-off equal to 100 were averaged over all images of the category) as in [34].

Examples of our retrieval results are shown in Fig. 7 and the *Average Precision* is given for each category in Fig. 8 (dark blue bars). The results of the *UFM* and *CLUE* approaches are also displayed in this latter figure for comparison. Fig. 7 illustrates the fact that the most of time, the first four retrieved images belong to the query's category (row 1, 4, and 5). This figure also illustrates well the difficulties encountered in this task: since the categories are quite large and diverse, images belonging to different categories may have very similar visual properties that are picked by our method. For example, the elephant and building (row 2 of Fig. 7) have dominating vertical structures and same dominant colors. Likewise, images belonging the "mountains" or "beaches" are freqently mismatched (row 3 of Fig. 7). These retrieval errors are common to all methods comparing images solely on the basis of the image content (i.e. introducing no semantics) and explain the fluctuation of the results displayed in Fig. 8 for all three methods. Our method compares well with the two established methods displayed here: it is more accurate than *UFM* (gray bars) for six categories out of ten; the accuracy is also better than or comparable to *CLUE* (white bars) for five categories out of ten. On average, our method performs better than the *UFM* approach and slightly less well than the *CLUE* one. As pointed out in Section 4.1, the *SMP* and *UFM* approaches are strictly content-based approaches. The *CLUE* method, while performing better, uses additional image distances and is therefore much more time-consuming. Thus, the performances of our method seem quite promising for three reasons:

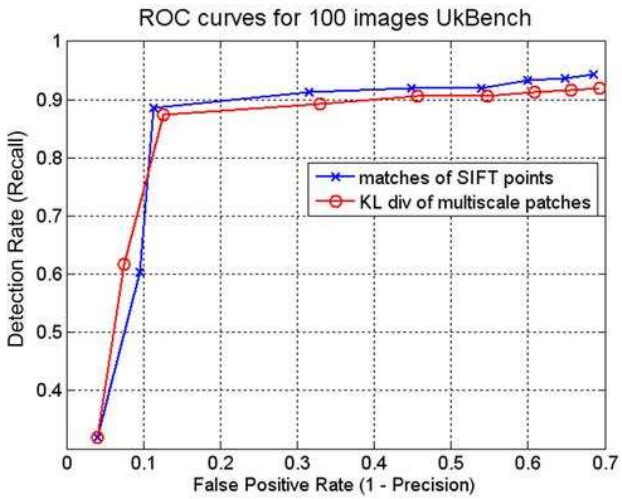

Fig. 6. Comparison of the retrieval performances of the *SMP* approach and the *SIFT*-based algorithm; the ROC curves were obtained for cut-off values going from 1 to 9.

- It performs slightly better than the *UFM* approach which relies on the same information.
- The results are not far from those of the more advanced *CLUE* approach which relies on more information.
- A similar clustering processing as the one applied with the *UFM* measure in *CLUE* may be applied to improve the *SMP* approach.

In conclusion, in its current state of development, the proposed *SMP* measure does not outperform the state-of-the-art methods selected as benchmark here. However, it does bring a novel approach to tackle the problem of image retrieval.

4.4 Computational speed-up(s)

The evaluation of our *SMP* similarity requires the computation of several KL divergences in a non-parametric framework. Since this is a time-consuming task, we propose two ways to speed-up the computations. The first one is based on a GPU implementation of the algorithm, the second on a preselection of the relevant images in the database.

GPU implementation

When computing the similarity between two images with the *SMP* approach, most of the time is devoted to the search of the k-th nearest neighbors in the evaluation of the KL divergences. Indeed, finding a k-th nearest neighbor requires to compute and sort distances between features (here the patches). The "brute force" algorithm has a complexity of order $O(N^2)$ for N samples in the feature

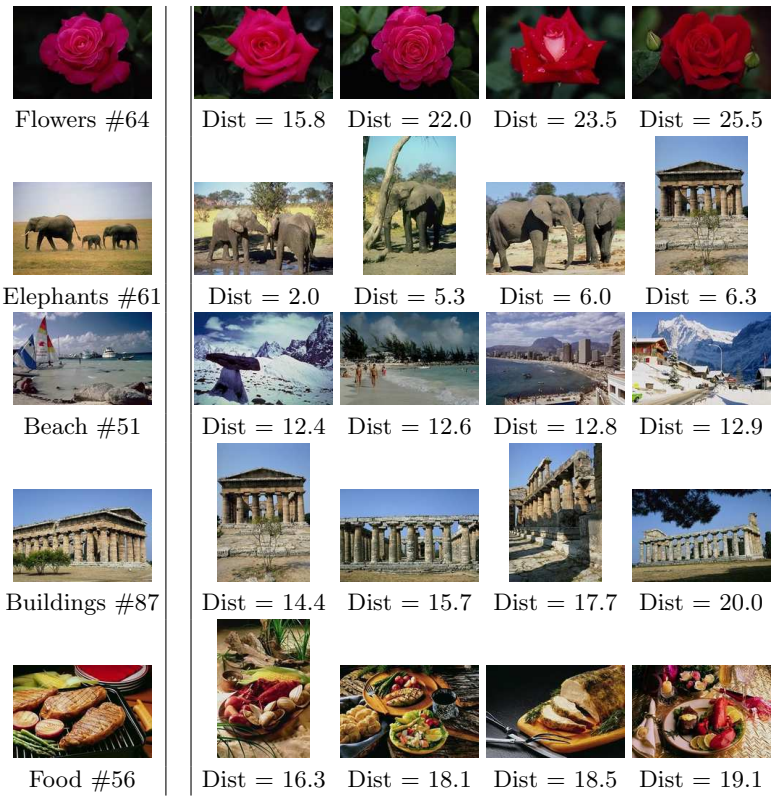

Fig. 7. Retrieval results for 5 images of the COREL database. For each row, left to right: query image; first 4 ranked images of the database (excluding the query image). For each retrieved image, the *SMP* similarity measure to the query is also shown.

set. Smarter algorithms with a lower complexity (typically of order $O(N \log N)$) such as the KD-tree-based (ANN) algorithm [36] have been designed. Nevertheless, in practice, the computation time of a similarity between two images with the *SMP* approach remains large even with this low-complexity algorithm: on average 2.2s on a Pentium 4 3.4 GHz (2GB of DDR memory) with the ANN algorithm.

To speed up the computation time, we developed a parallel implementation of the kNN search on a Graphic Processing Unit (GPU) [37] using CUDA. This implementation is based on a brute force approach since recursive algorithms (the preferred strategy when using trees such as in ANN) are not parallelizable. It was implemented on an NVIDIA GeForce 8800 GTX card with 768 MB of memory. The computation time for one similarity measure between two images required 0.2s on average (i.e., 10 times less than with the CPU implementation of ANN).

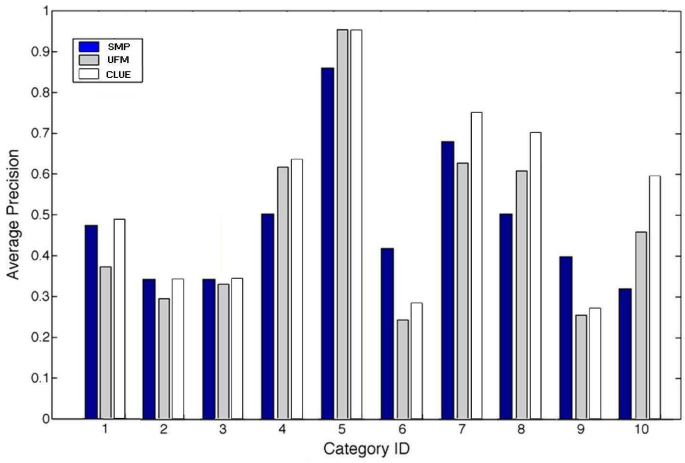

Fig. 8. Average Precision for each category of the COREL database. Dark blue bars: *SMP* approach; gray bars: *UFM* approach; white bars: *CLUE* approach.

As of today, the brute force algorithm parallelized on GPU is by far the fastest implementation of our method. Developing smart algorithms (such as the KD-tree one), which may not be parallelizable but have a very low complexity, is a topic of active research, as is the development of GPU for computational purposes. Hence both types of methods should be kept in mind for efficient implementations in the near future.

Preselection of the relevant images

The computational speed can be improved by splitting the retrieval task into two steps:

1. Only the low frequency contribution to the similarity measure defined in Eq. (4) is computed for all images in the database. This "partial" similarity measure produces a first ranking of the database images from which the first n images are selected for the next step.
2. The complete similarity measure is computed between the query and the n selected images.

This procedure saves computation time as it computes the whole similarity measure only for a reduced number of images (computing only part of it for images that are unlikely to be relevant to the query). The smaller the size of the preselected subset, the greater the improvement in terms of computation time. For example, when a query on the Nister database is processed following the described two-step procedure with a selected subset of 50 images, the average computation time per image drops from 0.2s to about 0.06s with the GPU implementation (and with similar retrieval performances). It is clear however that the number of preselected images cannot be arbitrarily small without seriously

affecting retrieval performances. It should be large enough compared to the number of images in the database as well as the number of relevant images for the query.

5 Conclusion

In this paper, we proposed a new image similarity framework based on high-dimensional probability distributions of patches of multiscale coefficients which we call *Sparse Multiscale Patches*. Feature sets are represented by these patches of subband coefficients that take into account intrascale, interscale and interchannel dependencies. The similarity between two images was defined as a linear combination of the "closeness" between the distributions of their features at each scale measured by the Kullback-Leibler divergence. The Kullback-Leibler divergences are estimated in a non-parametric framework, via a kNN approach. The proposed similarity measure seems to be stable when selecting a reduced number of patches, proving that a few significant patches are enough to represent the image features. This is a consequence of the sparsity of the multiscale transform.

We applied this framework to image retrieval. The proposed approach takes advantage of the properties of its global multiscale descriptors. In particular, it is robust to JPEG2000 compression (i.e. it matches the visual similarity between images with different amounts of blur or compression noise). Retrieval experiments were conducted on two publicly available datasets of real world images (Nister Recognition Benchmark and the COREL database) to evaluate the average performances of the method. In particular, the Nister dataset was used to benchmark the robustness to several geometric image deformations, such as change of viewpoint, rotation and zoom. Our results showed the reliability of the *SMP* approach with respect to these deformations. In addition, although our method is new, its performances tested on two databases are very close to those of several established retrieval methods: a reference retrieval method based on (local) *SIFT* descriptors and two versions of a fuzzy, segmentation-based *UFM* approach: *UFM* and *CLUE* . This indicates that the *SMP* approach adapts to quite different retrieval tasks, from the object level (on the Nister database) to the level of general categories (on the COREL database). Finally, our Sparse Multiscale Patches approach follows the same multiscale philosophy as the new compression standard SVC [38]. This presumes nearly straightforward use of low-level bitstream components in a foreseen extension of this method to video retrieval.

References

1. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: An experimental comparison. Information Retrieval **11** (2008) 77–107
2. Loupias, E., Sebe, N., Bres, S., Jolion, J.M.: Wavelet-based salient points for image retrieval. In: ICIP. Volume 2. (2000) 518–521

3. Swain, M., Ballard, D.: Color indexing. IJCV **7** (1991) 11–32
4. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27** (2005) 1615–1630
5. Huang, J., Mumford, D.: Statistics of natural images and models. In: CVPR. Volume 1., Fort Collins, CO, USA (1999) 541–547
6. Black, M., Anandan, P.: A framework for the robust estimation of optical flow. In: ICCV, Berlin, Germany (1993) 231–236
7. Black, M.J., Anandan, P.: The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. CVIU **63** (1996) 75–104
8. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: CVPR. Volume 2. (2000) 142–149
9. Viola, P., W.M. Wells, I.: Alignment by maximization of mutual information. IJCV **24** (1997) 137–154
10. Bansal, R., Staib, L.H., Chen, Z., Rangarajan, A., Knisely, J., Nath, R., Duncan, J.S.: Entropy-based, multiple-portal-to-3dct registration for prostate radiotherapy using iteratively estimated segmentation. In: MICCAI, London, UK (1999) 567–578
11. Marques, O., Mayron, L.M., Borba, G.B., Gamba, H.R.: On the potential of incorporating knowledge of human visual attention into cbir systems. In: ICME. (2006) 773–776
12. Puzicha, J., Rubner, Y., Tomasi, C., Buhmann, J.M.: Empirical evaluation of dissimilarity measures for color and texture. In: ICCV. (1999) 1165–1172
13. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Multiscale Modeling and Simulation **4** (2005) 490–530
14. Awate, S.P., Whitaker, R.T.: Unsupervised, information-theoretic, adaptive image filtering for image restoration. IEEE Trans. on PAMI **28** (2006) 364–376
15. Angelino, C.V., Debreuve, E., Barlaud, M.: Image restoration using a knn-variant of the mean-shift. In: ICIP, San Diego, USA (2008)
16. Portilla, J., Strela, V., Wainwright, M., Simoncelli, E.P.: Image denoising using a scale mixture of Gaussians in the wavelet domain. TIP **12** (2003) 1338–1351
17. Do, M., Vetterli, M.: Wavelet based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. TIP **11** (2002) 146–158
18. Wang, Z., Wu, G., Sheikh, H.R., Simoncelli, E.P., Yang, E.H., Bovik, A.C.: Quality-aware images. TIP **15** (2006) 1680–1689
19. Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81** (1994) 425–455
20. Romberg, J.K., Choi, H., Baraniuk, R.G.: Bayesian tree-structured image modeling using wavelet-domain hidden markov models. TIP **10** (2001) 1056–1068
21. Pierpaoli, E., Anthoine, S., Huffenberger, K., Daubechies, I.: Reconstructing sunyaev-zeldovich clusters in future cmb experiments. Mon. Not. Roy. Astron. Soc. **359** (2005) 261–271
22. Burt, P.J., Adelson, E.H.: The Laplacian pyramid as a compact image code. IEEE Trans. Communications **31** (1983) 532–540
23. Piro, P., Anthoine, S., Debreuve, E., Barlaud, M.: Image retrieval via kullback-leibler divergence of patches of multiscale coefficients in the knn framework. In: CBMI, London, UK (2008)
24. Nielsen, F., Boissonnat, J.D., Nock, R.: On bregman voronoi diagrams. In: SODA. (2007) 746–755
25. Nielsen, F., Nock, R.: On the smallest enclosing information disk. Inf. Process. Lett. **105** (2008) 93–97

26. Boltz, S., Debreuve, E., Barlaud, M.: High-dimensional kullback-leibler distance for region-of-interest tracking: Application to combining a soft geometric constraint with radiometry. In: CVPR, Minneapolis, USA (2007)
27. Ahmad, I., Lin, P.E.: A nonparametricestimationof the entropy absolutely continuousdistributions. IEEE Trans.Inform. Theory **22** (1976) 372–375
28. Terrell, George R.and Scott, D.W.: Variable kernel density estimation. The Annals of Statistics **20** (1992) 1236–1265
29. Loftsgaarden, D., Quesenberry, C.: A nonparametric estimate of a multivariate density function. AMS **36** (1965) 1049–1051
30. Goria, M., Leonenko, N., Mergel, V., Novi Inverardi, P.: A new class of random vector entropy estimators and its applications in testing statistical hypotheses. J. Nonparametr. Stat. **17** (2005) 277–298
31. Lowe, D.: Distinctive image features from scale-invariant keypoints. In: IJCV. Volume 20. (2003) 91–110
32. Chen, Y., Wang, J.Z.: A region-based fuzzy feature matching approach to content-based image retrieval. TIP **24** (2003) 1252–1267
33. Nistér, D., Stewénius, H.: Scalable recognition with a vocabulary tree. In: CVPR. Volume 2. (2006) 2161–2168
34. Chen, Y., Wang, J.Z., Krovetz, R.: Clue: Cluster-based retrieval of images by unsupervised learning. TIP **14** (2005) 1187–1201
35. Lowe, D.: Sift keypoint detector. (http://www.cs.ubc.ca/∼lowe/keypoints/)
36. Arya, S., Mount, D.M., Netanyahu, N.S., Silverman, R., Wu, A.Y.: An optimal algorithm for approximate nearest neighbor searching fixed dimensions. J. ACM **45** (1998) 891–923
37. Garcia, V., Debreuve, E., Barlaud, M.: Fast k nearest neighbor search using gpu. In: CVPR Workshop on Computer Vision on GPU. (2008)
38. ITU-T, JTC1, I.: Scalable video coding - joint draft 6 (Apr 2006)