Published in CBMI, 18-20/06/2008, London, UK - Copyright IEEE

IMAGE RETRIEVAL VIA KULLBACK-LEIBLER DIVERGENCE OF PATCHES OF MULTISCALE COEFFICIENTS IN THE KNN FRAMEWORK

Paolo Piro, Sandrine Anthoine, Eric Debreuve, Michel Barlaud

I3S lab., Université de Nice Sophia-Antipolis / CNRS; Sophia-Antipolis, France
{piro, anthoine, debreuve, barlaud}@i3s.unice.fr

ABSTRACT

In this paper, we define a similarity measure between images in the context of (indexing and) retrieval. We use the Kullback-Leibler (KL) divergence to compare sparse multiscale image representations. The KL divergence between parameterized marginal distributions of wavelet coefficients has already been used as a similarity measure between images. Here we use the Laplacian pyramid and consider the dependencies between coefficients by means of non parametric distributions of mixed intra/interscale and interchannel patches. To cope with the high-dimensionality of the resulting description space, we estimate the KL divergences in the k-th nearest neighbor (kNN) framework (instead of classical fixed size kernel methods). Query-by-example experiments show the accuracy and robustness of the method.

Index Terms— Image retrieval, sparse wavelet description, intra/interscale dependency, Kullback-Leibler divergence, k-th nearest neighbors.

1. INTRODUCTION

A central question in content-based image indexing is to define a similarity measure between images that matches - or at least is close enough to - our perception of their similarity. Then, database images can be simply ranked in increasing order of their similarity to the reference (or example) image for a query-by-example task. Understanding how human perceive the similarity between images via perceptual studies is still a topic of active research. Thus, content-based image indexing systems relying on such studies may be subjective and very hard to implement. Here, we focus on developing an objective and mathematically defined measure that will be easily implementable.

In this context, used measures often rely on global descriptions such as dominant colors or color distribution, or on extracted information such as salient points/regions together with local features or segmentation along with region arrangement [1, 2]. The philosophy here is to use a sparse multiscale image description. The Kullback-Leibler (KL) divergence has already been used as a similarity measure between parameterized marginal distributions of wavelet coefficients at different scales [3, 4]. Nevertheless, independence between the coefficients was assumed, preventing from taking into account local image structures such as texture. In contrast, we propose to consider dependency by means of distributions of mixed intra/interscale patches of the Laplacian pyramid coefficients. In addition, for the case of color images, we take into account the statistical dependencies amongst the three color channels; hence patches of coefficients are also interchannel. This approach implies to deal with a high-dimensional statistical description space. The number of samples being too small to reasonably fill this space, fixed size kernel options to estimate distributions or divergences fail. Alternatively, we propose to estimate the KL divergence in the k-th nearest neighbor (kNN) framework [5], *i.e.*, adapting to the local sample density and directly from the samples.

This work is supported in part by the ANR "ICOS-HD".

2. SIMILARITY BETWEEN IMAGES

In this section we define our similarity measure between images. It is a combination of Kullback-Leibler divergences in a multiscale feature space. This feature space consists of inter/intrascale and interchannel patches of Laplacian pyramid coefficients for color images. We first describe the feature space and then the use and estimation of the KL divergences.

2.1. Patches of Laplacian pyramid coefficients

Let us denote by $w(I)_{j,k}$ the coefficient for image I at scale j and location in space k for a general multiresolution decomposition.

The concept of patches of multiresolution coefficients was introduced by [6] for the wavelet decomposition under the name "neighborhoods of wavelet coefficients". It stems from the following ideas.

The wavelet domain provides a sparse representation of images, meaning that it concentrates the information into a few coefficients of large amplitude (the rest of the coefficients being small) and enjoys a fast transform. This is what makes wavelet thresholding methods such powerful tools in image processing. Initial thresholding wavelet methods treat each

coefficient separately relying on the fact that these coefficients are decorrelated. However, they 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 ($w(I)_{j,k_0}$ large for all j). The most significant dependencies are seen between a wavelet coefficient $w(I)_{j,k}$ and its closest neighbors in scale: $w(I)_{j-1,k}$, or in space: $w(I)_{j,k\pm(0,1)}$, $w(I)_{j,k\pm(1,0)}$. Several models using these dependencies have been proposed and used in image enhancement [6, 7]. The concept of patches (or neighborhoods) of wavelet coefficients, was introduced in [6]. These are vectors of the form:

$$\mathbf{w}(I)_{j,k} = \left(w(I)_{j,k}, w(I)_{j,k\pm(1,0)}, w(I)_{j,k\pm(0,1)}, w(I)_{j-1,k}\right) \quad (1)$$

for a greyscale image.

The probability density function (pdf) of such patches was shown to characterize and estimate fine spatial structures in greyscale images [6, 8]. Hence such patches are expected to be relevant features to represent the image content.

Critically sampled tensor wavelet transforms lack of rotation and translation invariance and so would the neighborhoods made of such coefficients. We prefer to use the Laplacian pyramid [9] which shares with the wavelet transform the sparsity and inter/intrascale dependencies properties and is more robust to rotations. Hence we will define our feature space from the set of patches of Laplacian pyramid coefficients of the form of (1).

Here, we consider colored images in the luminance - chrominances space: $I = (I^Y, I^U, I^V)$. Since the pyramid coefficients are correlated through channels, we aggregate in the patches the coefficients of the three channels (still noting the patches $\mathbf{w}$):

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

with

$$\mathbf{w}(I^c)_{j,k} = \left(w(I^c)_{j,k}, w(I^c)_{j,k\pm(1,0)}, w(I^c)_{j,k\pm(0,1)}, w(I^c)_{j-1,k}\right). \quad (3)$$

Hence our feature space is the set of the patches as in Eq. (2) for all scales j and locations k. Fig. 1 illustrates how to build one patch from two near subbands of the Laplacian pyramid in the case of a single color component image (as in Eq. (1)).

The low-frequency subband is also considered in a similar way. Namely, intrascale and interchannel patches (of dimension 27) are built by joining the spatial 3×3 neighborhoods in the three channels. As a result, our feature space is the set of Laplacian pyramid coefficients defined in Eq. (2) (of dimension 18), for all scales j and locations k in addition to the low-frequency patches.

Let us now turn to the measure of similarity on this space.

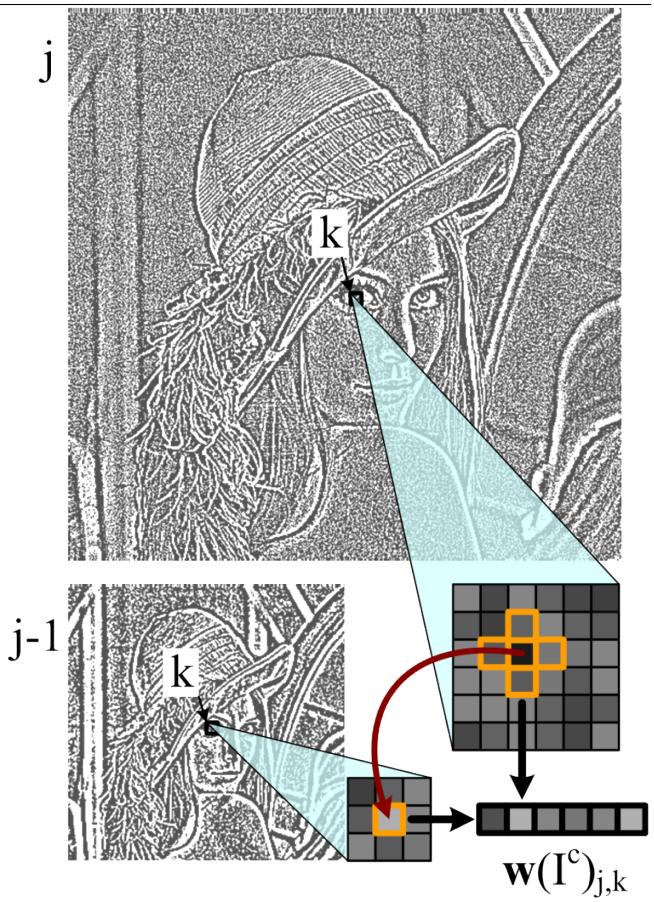

Fig. 1. How to build a patch by grouping multiscale coefficients, for a single color channel.

2.2. Similarity measure between images

Since geometrically modified or slightly degraded versions of the same image as well as images containing similar objects should be close, one cannot define a measure comparing directly the neighborhoods one by one, but rather their probability distributions. More specifically, we consider the pdf of the neighborhoods of Eq. (2) for each scale, i.e. we consider the pdf $p_{\mathbf{w}_j(I)}$ of the set neighborhoods $\{\mathbf{w}(I)_{j,k}\}_k$ for each fixed j.

The considered pdf are those of coefficients that carry the informational content on the signal when they are large. The natural way to compare such pdf is to use measures derived from information theory. Here, we use the KL divergence between pdf, an approach that has also been successfully taken for other applications [5, 10]. This was also done in [3, 4] to evaluate the similarity between images using the marginal pdf of the wavelet coefficients. We propose to use this measure on the multidimensional pdf of the neighborhoods of coefficients: the similarity between images I_1 and I_2 is a weighted sum over scales of the KL divergences between the

pdf $p_{\mathbf{w}_j(I_1)}$ and $p_{\mathbf{w}_j(I_2)}$:

$$S(I_1, I_2) = \sum_j \alpha_j D_{kl}(p_{\mathbf{w}_j(I_1)} || p_{\mathbf{w}_j(I_2)}) \quad (4)$$

where $p_{\mathbf{w}_j(I_i)}$ is the pdf of the pyramid patches (Eq.(2)) of image I_i at scale j and $\alpha_j > 0$ are weights (chosen according to the redundancy of the transform in each scale).

Previous works on neighborhoods of wavelet coefficients or indexation using marginal pdf of the wavelet coefficients all assumed a parametric model for the pdf involved. In the marginal case, efficient models (e.g. generalized Gaussian [3, 4]) lead to an analytic expression of the KL divergence as a function of the model parameters; but they are not easily generalizable to the multidimensional correlated case of multiscale patches. On the other hand, efficients multidimensional models including correlations (e.g. Gaussian mixtures [8]) fit a wide variety of multidimensional pdf but impose to estimate the KL divergence after estimating the model parameters. Besides the heavy computational cost of the consecutive estimations, the stability of such cascading estimates is likely to be difficult to obtain numerically. We prefer to make no hypothesis on the pdf at hand, hence sparing the cost of fitting the model parameters but with the need of estimating the KL divergences in this non-parametric case.

2.3. Estimation of the Kullback-Leibler (KL) divergence

Let us first remind that the expression of the KL divergence between two continuous pdf p_1 and p_2 is:

$$D_{kl}(p_1||p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx = H_x(p_1, p_2) - H(p_1) \quad (5)$$

with H the differential entropy and H_x the cross entropy.

The estimation of statistical measures in the multidimensional case is hard. In particular, kernel-based methods such as Parzen estimates become unadapted due to the sparsity of samples in high dimension (curse of dimensionality): the tradeoff between a kernel with a large bandwidth to perform well in low local sample density (which *oversmoothes* the estimator) and a kernel with a smaller bandwidth to preserve local statistical variability (which results in an unstable estimator) cannot always be achieved. We use instead the k-th nearest neighbor (kNN) framework [11] to compute the KL divergence. Indeed it follows the dual approach to the above fixed size kernel: the bandwidth adapts to the local sample density by letting the kernel contain exactly k neighbors of a given sample. Moreover it allows direct estimation of the divergence without explicitly estimating the pdf.

Fix ν, a set of N_ν samples $\mathbf{w}_1, \mathbf{w}_2, .., \mathbf{w}_{N_\nu}$ of pdf p_ν and k, a non-negative integer. kNN balloon estimates are based on the principle that $p_\nu(s)$ is inversely proportional to the volume of the sphere containing the k nearest neighbors of s in ν [11]:

$$p_\nu(s) \sim \frac{k}{v_d \, \rho_{k,\nu}^d(s)} \quad (6)$$

with v_d the volume of the unit sphere in $\mathbb{R}^d$ and $\rho_{k,\nu}(s)$ the distance of s to its k-th nearest neighbor in $\nu - \{s\}$.

Plugging this density estimator in Ahmad-Lin [12] entropy estimator:

$$H_{al}(p_\nu) = -\frac{1}{N_\nu} \sum_{n=1}^{N_\nu} \log\Big[p_\nu(\mathbf{w}_n)\Big] \quad (7)$$

we get a biased entropy estimate. This bias can be rid of by replacing $\log(k)$ by the digamma function $\psi(k)$, giving the entropy estimate [13]:

$$\widehat{H}(p_\nu) = \log\Big[(N_\nu - 1) v_d\Big] - \psi(k) + \frac{d}{N_\nu} \sum_{n=1}^{N_\nu} \log\Big[\rho_{k,\nu}(\mathbf{w}_n)\Big] \quad (8)$$

The cross entropy estimate is then [5]:

$$\widehat{H_x}(p_{\nu_1}, p_{\nu_2}) = \log\Big[N_{\nu_2} v_d\Big] - \psi(k) + \frac{d}{N_{\nu_2}} \sum_{n=1}^{N_{\nu_2}} \log\Big[\rho_{k,\nu_1}(\mathbf{w}_n^2)\Big] \quad (9)$$

and the KL divergence estimate is:

$$\begin{aligned} \widehat{D_{\mathrm{kl}}}(p_{\nu_1}||p_{\nu_2}) &= \log\Big[\tfrac{N_{\nu_2}}{N_{\nu_1}-1}\Big] + \frac{d}{N_{\nu_2}} \sum_{n=1}^{N_{\nu_2}} \log\Big[\rho_{k,\nu_1}(\mathbf{w}_n^2)\Big] \\ &\quad - \frac{d}{N_{\nu_1}} \sum_{n=1}^{N_{\nu_1}} \log\Big[\rho_{k,\nu_1}(\mathbf{w}_n^1)\Big] \end{aligned} \quad (10)$$

This expression is valid in any dimension and it is robust to the choice of k.

3. NUMERICAL EXPERIMENTS

3.1. Settings

We used for our numerical experiments a database containing images from the Recognition Benchmark collection, which is a ground-truth dataset already used in [14]. The database consists in 640x480 images grouped by sets of four images of the same scene. Hence, for any query image, exactly three other relevant images are to be retrieved. A subset of 100 images from this dataset was used to adjust some parameters of our algorithm and compare its performance with a reference method (section 3.2 and 3.5). The method we propose was then validated on a larger dataset, containing 1,000 images from the same collection (section 3.3 and 3.4).

In all experiments a Laplacian pyramid was computed for each channel of the images (in the YUV color space) with a 5-points binomial filter. The first three high-frequency subbands and the low-frequency image approximation were used to build patches.

To perform the KL divergence estimations, we used a fast kNN search algorithm of complexity $O(N \log N)$, N being the number of sample points [15]. This algorithm was run on

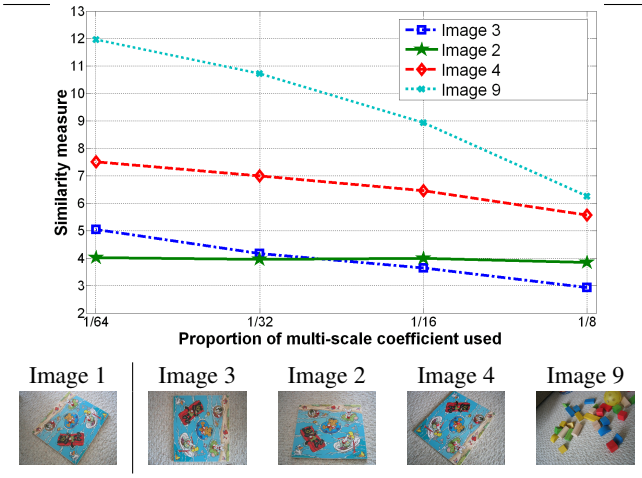

Fig. 2. Evolution of the similarity measure with the proportion of patches selected. Retrieval results shown for image query 1 and its four most similar images (images 2, 3 and 4 are relevant for this query).

a CPU Pentium 4 3.4 GHz with 2GB of DDR memory. The average time required to compute the distance between two images was about 2.2 s.

In order to speed up the computing time for queries on a large dataset, we used a parallel implementation of the kNN searching on Graphic Processing Unit (GPU) [16]. This is based on a brute-force approach and was written in the CUDA development environment. It was run on a NVIDIA GeForce 8800 GTX graphic card and allowed to compute a single similarity measure in less than 0.2 s on average.

The sparsity property of the multiresolution transform allowed us to retain not all feature vectors but only a small proportion of them in the high-frequency subbands, thus reducing the number of operations required for kNN estimations. For this purpose, only a certain proportion of patches was taken into account, according to a significance criterion. Namely, all subbands in the luminance channel were thresholded based on either the energy (quadratic norm) of the whole patch or the amplitude of the patch center, so that all coefficients being under the threshold do not contribute to entropy estimation. Then, each significant patch in the luminance channel and the corresponding ones in the two chrominance channels were joined together as described in section 2.1.

With respect to the kNN estimation of the KL divergence, we fixed k to ten and the weights α_j of Eq.(4) to one, so that all subbands contribute equally to the similarity measure.

3.2. Robustness to sparsity of multiscale features

The similarity measure we propose is robust with respect to the sparsity of the subband image decomposition. Indeed, retrieval results are consistent with ground-truth, as we decrease the number of subband patches contributing to distance estimations. In order to select the most significant patches, (i.e. those containing the most important information of the image), we selected the largest ones (either in the sense of overall energy or in the sense of center amplitude). Given a subset of the database containing both relevant and non relevant images for a given query, we computed the similarity measures using different proportions of significant patches. Results are generally consistent as sparsity increases (i.e. as the number of selected patches decreases), in the sense that relevant images remain significantly closer to the query than others and the rank order of retrieved images does not change. An example is given in Fig. 2, where similarity between one query and the first four retrieved images is shown for different proportions of selected subband coefficients. This trend is general for all pictures and confirms the suitability of the selection criterion: it allows to reduce the computing time for the distance estimation, while preserving its accuracy and sensitivity.

3.3. Retrieval results

Results for five queries in the database containing 1,000 images are displayed in Fig. 3. In this figure, each row displays the retrieval result for the query image shown on the leftmost column. From the second column on, one can see the first 3 retrieved images ranked in increasing order of their similarity measure from the query. Hence the second leftmost image is the most similar, excluding the reference image which is always ranked first with a distance of zero. The criterion we adopted to select feature vectors was based on patch energy. Proportions of selected patches in the three considered high-frequency subbands were respectively: $1/64$, $1/32$, $1/16$.

As shown by experimental results, the method we propose seems to perform very well. In particular, images being relevant for a given query are generally ranked first; this holds in spite of the fact that images belonging to the same group have been often subjected to different geometric transformations. Hence our method is robust to such transformations, e.g. rotations (such as the third row in Fig. 3) or changes in viewpoint (rows 2 and 5) and zoom (rows 1 and 4).

3.4. Evaluation of image retrieval performances

A wide variety of measures have been proposed to evaluate image retrieval performance [17]. We adopt a standard criterion in information retrieval, which has been also used in the context of image retrieval [18]. It is based on the number R of expected results for a given query (*relevant images*), the number D of correct results (*detected images*) and the number W of wrong results (*false positive images*), with respect to the ground-truth relevance. Naturally, these measures depend on the number C of retrieved images (*cut-off*). Together, they enable to define the standard measures of *precision* and *recall* in the following way:

$$precision = \frac{D}{C} = \frac{D}{D+W}, \quad recall = \frac{D}{R}\,. \qquad (11)$$

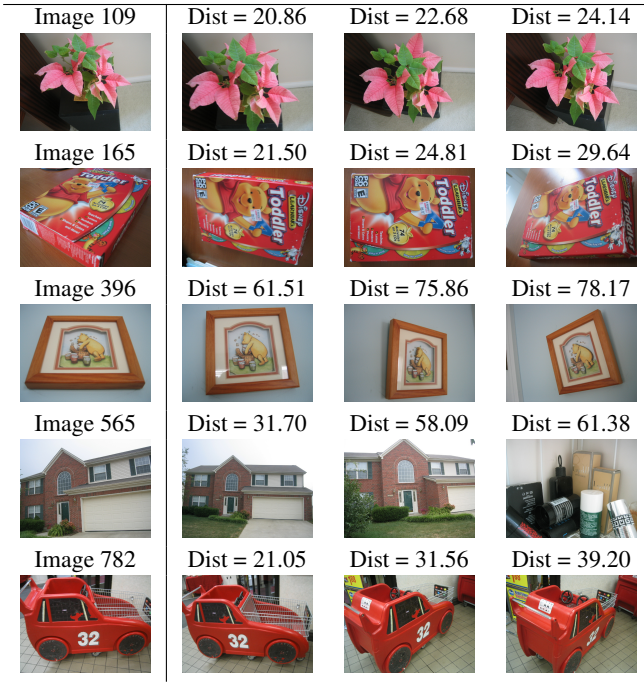

Fig. 3. Retrieval results for 5 images of the database. For each row, left to right: reference image; first 3 ranked images of the database (excluding reference image). For each retrieved image our distance to the query is also shown.

Varying the cut-off value, one obtains the ROC (Receiver Operating Characteristic) curve, which gives the performance of the retrieval system for each request. This curve is generally shown as $recall$ versus $1 - precision$, that represent, respectively, the detection rate and the false positive rate. Hence the larger are precision and recall values, the better is retrieval performance.

To evaluate the overall performance of our method, we used the curves obtained by querying the whole database with each of the 1,000 images. The significance criterion for patch selection was that based on patch energy, that showed the best performance.

By averaging over all queries the individual values of precision and recall, we obtained one average curve. It is displayed in Fig. 4. This graph shows that the best trade-off between precision and recall was reached when we retrieved three images; in this case, cut-off value matches exactly the number of relevant images or, in other words, there is a high probability that retrieved images are all and only the relevant ones.

3.5. Comparison with SIFT-based retrieval

Retrieval experiments were done by using a state-of-the-art method as well; this is based on SIFT descriptors, that represent the gradient orientations at interest points. It uses a

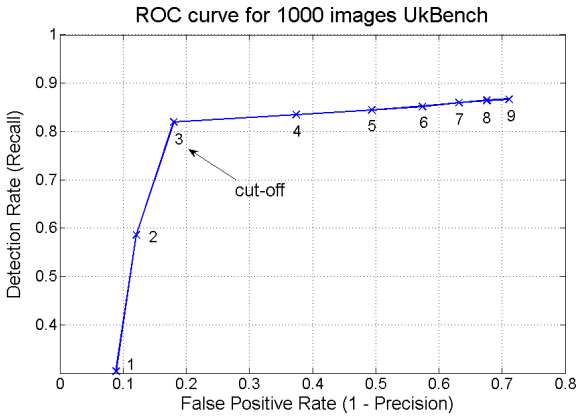

Fig. 4. Receiver Operating Characteristic (ROC) from cut-off variation.

matching criterion between the interest points of images to measure their similarity. We used a Matlab implementation of the SIFT algorithm [19]. Since it took about 4.8 s on average to compute the similarity between two images, experiments on the whole database could not be performed in a reasonable time. As a result, we compared the two methods by querying a subset containing 100 images. The results are shown in Fig. 5 in terms of the ROC curves. Namely, the KL divergence between patch distributions and the number of matched SIFT keypoints between images are compared as similarity measures for image retrieval.

The performances of our method are shown to be very close to that of the SIFT-based method, which is a reference method for the task of content-based image retrieval.

4. CONCLUSION

In this paper, we proposed a new image retrieval method based on multidimensional probability distributions of multiscale coefficients. These are grouped in coherent patches, that are selected by a significance criterion in order to effectively represent image features. The patches are build by taking into account intrascale, interscale and interchannel dependencies of subband coefficients for color images. The similarity measure we used is the sum over scales of the KL divergences between probability distributions of the image features, and it is estimated non-parametrically via a kNN approach. It was proven to be consistent with respect to the sparsity of the subband transform.

Query-by-example experiments on real images confirm the suitability of the proposed measure in the context of image retrieval. It is in particular robust to different geometric transformations, such as change in viewpoint, rotation and zoom. Moreover, retrieval performances are comparable to those of a reference algorithm, based on local SIFT descriptors.

Further investigations are planned with larger databases

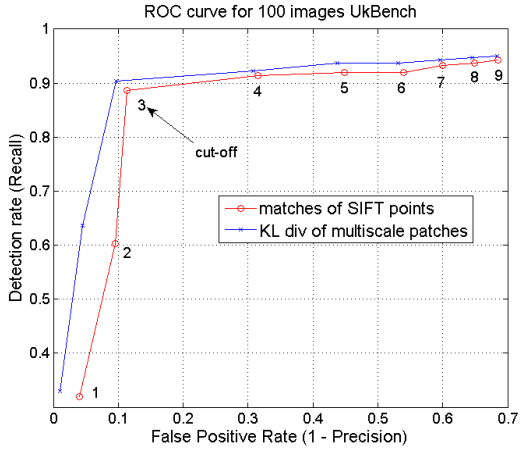

Fig. 5. Performance of image retrieval for our method and SIFT-based algorithm in terms of the ROC curve.

and others benchmark image collections. Future works will also include a more extensive comparison with state-of-the-art CBIR systems and the extension of the proposed method to the retrieval of video sequences.

5. REFERENCES

[1] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Object-based mpeg-2 video indexing and retrieval in a collaborative environment," *Multimed. Tools Appl.*, vol. 30, pp. 255–272, 2006.

[2] Q. Zhang and E. Izquierdo, "Optimizing metrics combining low-level visual descriptors for image annotation and retrieval," in *ICASSP, Toulouse, France*, 2006.

[3] MN Do and M Vetterli, "Wavelet based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *IEEE TIP*, vol. 11, pp. 146–158, 2002.

[4] Z Wang, G Wu, H R Sheikh, E P Simoncelli, E-H Yang, and A C Bovik, "Quality-aware images.," *IEEE Trans. Image Process.*, vol. 15, pp. 1680–1689, 2006.

[5] S. Boltz, E. Debreuve, and M. Barlaud, "High-dimensional kullback-leibler distance for region-of-interest tracking: Application to combining a soft geometric constraint with radiometry," in *CVPR*, Minneapolis, USA, 2007.

[6] J Portilla, V Strela, M Wainwright, and E P Simoncelli, "Image denoising using a scale mixture of Gaussians in the wavelet domain," *IEEE Trans. Image Process.*, vol. 12, pp. 1338–1351, 2003.

[7] J K Romberg, H Choi, and R G Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain hidden markov models.," *IEEE Trans. Image Process.*, vol. 10, pp. 1056–1068, 2001.

[8] E. Pierpaoli, S. Anthoine, K. Huffenberger, and I. Daubechies, "Reconstructing sunyaev-zeldovich clusters in future cmb experiments," *Mon. Not. Roy. Astron. Soc.*, vol. 359, pp. 261–271, 2005.

[9] Peter J. Burt and Edward H. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. COM-31,4, pp. 532–540, 1983.

[10] C. V. Angelino, E. Debreuve, and M. Barlaud, "A nonparametric minimum entropy image deblurring algorithm," in *ICASSP*, Las Vegas, Nevada, U.S.A., April 2008.

[11] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, 1992.

[12] I. Ahmad and P-E. Lin, "A nonparametric estimation of the entropy absolutely continuous distributions," *IEEE Trans. Inform. Theory*, vol. 22, pp. 372–375, 1976.

[13] M Goria, N Leonenko, V Mergel, and P Novi Inverardi, "A new class of random vector entropy estimators and its applications in testing statistical hypotheses," *J. Nonparametr. Stat.*, vol. 17, pp. 277–298, 2005.

[14] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2006, vol. 2, pp. 2161–2168.

[15] Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *J. ACM*, vol. 45, no. 6, pp. 891–923, 1998.

[16] V. Garcia, E. Debreuve, and M. Barlaud, "Fast k Nearest Neighbor Search usign GPU," Tech. Rep. arXiv:0804.1448, I3S Laboratory, University of Nice Sophia-Antipolis / CNRS, Apr. 2008.

[17] J. Smith, "Image retrieval evaluation," in *IEEE Workshop on Content-based Access of Image and Video Libraries, Santa Barbara, California*, 1998.

[18] Krystian Mikolajczyk and Cordelia Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, 2005.

[19] D. Lowe, "Sift keypoint detector," http://www.cs.ubc.ca/∼lowe/keypoints/.