Combining spatial and temporal patches for scalable video indexing

Paolo Piro, Sandrine Anthoine, Eric Debreuve, Michel Barlaud

Multimedia Tools and Applications (Special Issue on CBMI2009.) manuscript No.
(will be inserted by the editor)

Combining spatial and temporal patches for scalable video indexing

P. Piro · S. Anthoine · E. Debreuve · M. Barlaud

Received: date / Accepted: date

Abstract This paper tackles the problem of scalable video indexing. We propose a new framework combining spatial and motion patch descriptors. The spatial descriptors are based on a multiscale description of the image and are called *Sparse Multiscale Patches*. We propose motion patch descriptors based on block motion that describe the motion in a Group of Pictures. The distributions of these sets of patches are compared combining weighted Kullback-Leibler divergences between spatial and motion patches. These divergences are estimated in a non-parametric framework using a k-th Nearest Neighbor estimator. We evaluate this weighted dissimilarity measure on selected videos from the ICOS-HD ANR project. Experiments show that the spatial part of the measure is relevant to detect different sequences, while its motion part allows to detect clips within a sequence. Experiments combining the spatial and temporal parts of the dissimilarity measure show its robustness to resampling and compression; thus exhibiting the spatial scalability of the method on heterogeneous networks.

Keywords Scalable video indexing · sparse multiscale patches descriptors · motion patches descriptors · Kullback-Leibler divergence

1 Introduction

In the last decades, the amount of video documents stored in databases has rapidly grown, together with the need for efficient tools to order, explore and use such databases. In addition, video documents, which generally show a large variety of size and formats, are to be available for retrieval through heterogeneous networks. Such networks connect different devices and technologies that are able to access to the video content with different performances (e.g., in terms of quality and resolution). Hence, defining a video indexing framework that is aware of retrieval capabilities of these different devices is a major challenge for content-based video retrieval systems. Based on these motivations, we investigate in this paper scalable descriptors for video indexing and propose a new framework for similarity-based video retrieval. We assume that no manual annotation of video documents is available, thus constraining the video search engine to handle content-based indices.

Several approaches have been proposed in the recent literature to tackle this problem. A first category of content based video indexing methods developed recently mainly uses global features of the video content such as the dominant color in still images or video key-frames [12]. These methods do not explicitly take into account the motion present in a video, and thus are not suitable to queries regarding the motion in a sequence (e.g. the task of finding videos with object having similar trajectories). Other methods explicitly take into account the

This work is supported by the French ANR grant “ICOS-HD”.

P. Piro · S. Anthoine · E. Debreuve · M. Barlaud
I3S lab., Université de Nice Sophia-Antipolis / CNRS
2000, route des Lucioles - Les Algorithmes - bât. Euclide B - BP 121 - 06903 Sophia Antipolis Cedex - France
Tel.: +33-492942749
Fax: +33-492942898
E-mail: {piro, anthoine, debreuve, barlaud}@i3s.unice.fr

motion and visual information in the video. Amongst these are object based video indexing methods [13,6,9, 5,3] that rely on a segmentation of the semantic objects in the video. The object is usually segmented spatially (except in [5] where foreground objects are segmented using the motion of MPEG-2 macroblocs) and the object motion is followed through the video. This spatio-temporal task is difficult to achieve since the objects undergo various transformations or may be occulted through the sequence.

Our objective in this paper is to provide a framework that will enable 1) to answer to different search tasks on video databases (e.g. find videos with similar motions or videos containing similar objects) and 2) to provide coherent answers with the various data formats that are available to the user through a heterogeneous network (scalability). I.e. we assume that the video data are available through a heterogeneous network. In this case, the end-user may have the video content in various formats according to his own device (PDA, desktop computer, etc.). The *Scalable Video Coding* (SVC) [10] standard in particular allows this variability in the network. We intend to design a method that is scalable in the sense that it gives similar answers whatever the format is that the end-users uses. Note that here, we focus on spatial scalability: we show that the proposed method gives a similar answer whatever the spatial format is that the user owns (the temporal scalability is not taken care of). To do so, we define a statistical dissimilarity measure between *Groups of Pictures* (GoPs) that is based on a complete spatio-temporal description of the video. We define two kinds of descriptors, 1) spatial descriptors that capture the visual content of a scene in a multiscale fashion and 2) temporal or motion descriptors that capture the motion in a GoP at the level of the block. Both kinds of descriptors are patch descriptors that exploit the spatial or temporal coherence present in the video. The sets of descriptors are compared statistically by a dissimilarity measure so that loose transformations of the video are not penalized (e.g. geometric or radiometric transformations, compression, etc.). We test this method on selected videos from the ICOS-HD ANR corpus that is designed specifically to probe the scalability of methods comparing videos and their robustness to radiometric and geometric transforms. We study the influence of the spatial and temporal parts of the proposed dissimilarity both separately and jointly.

The rest of the paper is organized as follows. Section 2 describes the descriptors we propose (both spatial and temporal). The dissimilarity measure is defined in Section 3, its practical implementation using the k-th nearest neighbors approach is detailed, and its scalability is discussed. Experiments showing the influence of both the temporal aspect and the spatial aspect of the proposed measure are given in Section 4, showing in particular its scalability on heterogeneous networks.

2 Spatio-temporal descriptors

We have previously developed spatial descriptors called *sparse multiscale patches (SMP)* and showed that they characterize the visual features of still images (see [7,8]). These descriptors provide a sparse description of the features of an image by grouping the coefficients of its multiscale transform into patches. To accurately describe videos, we also need descriptors of the apparent motion of the objects in the scene. We generalize the concept of *SMP*s to obtain descriptors of the apparent motion in each GoP of a video sequence.

2.1 Sparse multiscale patches (*SMP*)

Here is a brief review of our spatial descriptors described in details in [7,8]. To capture the local structure of an image at a given scale and at a specific location, we use a multiscale transform such as wavelet transform or Gabor transform to represent the image. We then form a patch of the *sparse multiscale patches* or *SMP* description by grouping multiscale coefficients of all color channels of the image that are neighbors across scale and location. More precisely, we note $w^{I^c}_{j,k}$ the multiscale coefficient of channel c of image I at scale j and location k (this would be the dot product of channel c of image I with a waveform of scale j centered at location k). We firstly group the coefficients of closest scale and location for each color channel to form the intermediate patches $\mathbf{w}^{I^c}_{j,k}$ (see Fig. 1):

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

(Note that scale $j-1$ is a coarser than scale j.)
To take into account the coherence of the local structures of image through color channels, interchannel patches

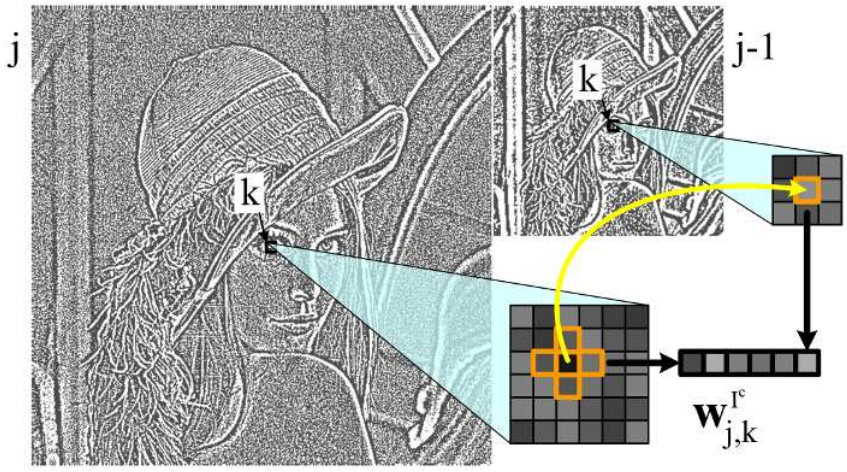

Fig. 1 Building a patch of multiscale coefficients, for a single color channel image.

$\mathbf{W}^I_{j,k}$ for color images in the YUV space are then formed grouping the patches of the three channels $\mathbf{w}^{I^Y}_{j,k}, \mathbf{w}^{I^U}_{j,k}$, and $\mathbf{w}^{I^V}_{j,k}$:

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

A single patch $\mathbf{W}^I_{j,k}$ captures the inter/intrascale and interchannel dependencies between neighboring multiscale coefficients which are the signature of local structures in the image. We use the Laplacian pyramid as a multiscale transform for its near-invariance properties towards rotations and translations and its reduced redundancy. Each patch $\mathbf{W}^I_{j,k}$ has length 18. The picture would not be complete without a description of the low frequency part of the image (the patches of Eq.(2) are built exclusively from the band-pass and high-pass subbands). Low-frequency patches are the concatenation across channels of 3 by 3 neighborhoods of the low-frequency coefficients of each color channel (making patches of length 27). To simplify the notation, let us from now on, denote by $\mathbf{W}^I_{j,k}$ either a low-pass or a high-pass or band-pass patch.

The sparsity properties of the multiscale transform transfer to the description by multiscale patches. Indeed, 1) the set of patches of large energy (sum of squared coefficients) is a small - or sparse - subset of the large set of all multiscale patches $\{\mathbf{W}^I_{j,k}\}_{j\geq j_0, k\in\mathbb{Z}}$ and 2) this small subset describes well the content of the image (this is a sparsity property: a small group yields a good representation). We select the so-called *sparse multiscale patches* by thresholding the energy level at each scale j and thus obtain our spatial descriptors of an image i.e. a frame of a video. (For example, for videos in HD format as in Section 4, the images are decomposed on five scales of the Laplacian pyramid and $1/6$ of the patches are kept at each scales, except for the lowest ones where all patches are used). The cost of the extraction of the *SMP* scales as $n \log n$ where n is the number of pixels in the image.

2.2 GoP motion patches (*GoP-MP*)

In this section, we present new temporal patch descriptors. To describe the motion in a video, we also use the concept of patches. Here, the patches are understood as groups of motion vectors that behave coherently. Since objects have naturally relatively smooth motion trajectories across restricted periods of time, the coherence is sought through time. The idea is to group in a patch the motion vectors that describe a coherent motion through the GoP. In video standards such as the *Scalable Video Coding* (SVC) standard [10], the motion information is encoded by coding macroblock motion. A first approach inspired by the compression standards is to encode block motion in the temporal patches. This can be done in two ways:

- Either one fixes the location of the block to a point (x,y) and estimates the motion from this point for each pair of frames of the GoP.
- Or one follows the trajectory of the block which is located at (x,y) in the first frame of the GoP.

Other approaches are possible, where the motion of the video is considered at a different spatial level than the block. For example:

- Assuming we have a coarse segmentation of the scene into its different objects and an estimation of the apparent motion of each of them; we could build a motion patch for each object that follows its trajectory through the GoP. The patch would then be the concatenation of the successive displacements of the object.
- We could also consider computing the optical flow between each successive frame to obtain the apparent motion of each single pixel through time. In this case, we would obtain a patch that follows the motion of this pixel.

Obtaining a coarse spatio-temporal segmentation of moving objects in a video or the optical flow is more complex and computationally more intensive than the block motion solutions. Moreover, we are here seeking a sparse representation of the GoP motion, therefore encoding the apparent motion of all pixels through the optical flow is not appropriate. Thus, in this paper we focus on motion patches at the block level, inspired by compression standards. Note that we consider here that blocks are all of the same size through the GoP (which may not be the case in compression standards). Moreover, in the set of experiments presented in Section 4, we keep the block location fixed through the GoP (experiments following block trajectories are not reported here as they give similar results).

The motion patches are computed as follows. We compute the apparent motion of each particular block (x,y) of a GoP (of around 8 to 10 pictures). More precisely, for a GoP of n consecutive frames $f_1, \ldots, f_n$, we compute the following motion patches for each block of center (x,y):

$$m(\mathrm{x,y}) = \left(x, y, \mathbf{u}_{1,2}(\mathrm{x,y}), \mathbf{u}_{2,3}(\mathrm{x,y}), \ldots, \mathbf{u}_{n-1,n}(\mathrm{x,y})\right) \tag{3}$$

where $\mathbf{u}_{n-1,n}(\mathrm{x,y})$ is the apparent motion of block (x,y) from frame f_{n-1} to frame f_n (see Fig. 2). Note that we include in the motion patch its location (x,y) so that each patch has length $2n+2$ (which is 18 for GoPs of 8 frames). This localization of the motion patches reflects the geometry of the underlying objects. We will exploit this property to compare sets of motion patches when defining our dissimilarity measure in the next section.

The motion vectors $\mathbf{u}$ are computed via a diamond-search block matching algorithm. For each GoP studied, we compute the motion patches m(x,y) for each block (x,y). As is the case for spatial patches, in fact only a few motion patches effectively describe motion (sparsity). Thus, we select the significant motion patches by a thresholding that keeps only the patches having the largest motion amplitude (sum of square of the $\mathbf{u}$ components in Eq. (3)).(For example, for videos in HD format as in Section 4, the motion patches kept are those for which the motion amplitude is non-zero). The cost of the extraction of the motion patches is the cost of performing the diamond-search block matching algorithm on the lowest scale of the Laplacian pyramid decomposition. Sequences longer than a GoP are divided in GoPs from which we extract the significant motion patches.

3 Measuring the dissimilarity between videos

Since the natural unit of time of our temporal descriptors is the GoP, we define a dissimilarity measure that compares GoPs on the basis of both temporal and spatial descriptors. To compare longer sequences such as clips, we simply add up the dissimilarity measures between their consecutive GoPs.

3.1 Comparing two GoPs

For a single GoP G, we consider both temporal and spatial descriptors. The set of temporal descriptors called M^G is selected as in Section 2.2. To represent the spatial information in a GoP of a video, we use the spatial descriptors of its first frame (this is sufficient since a GoP has a short duration). These are furthermore divided into several sets, more exactly, we group the *SMP*s $\mathbf{W}^G_{j,k}$ according to their scale index j. We obtain a set of *SMP*s noted $\mathbf{W}^G_j$ for each scale j of the multiscale decomposition.

We intend to define a dissimilarity that is scalable (in the sense that it adapts to the different formats available on heterogeneous networks) and robust to geometric deformations. Hence, given a query GoP G_q and a reference GoP G_r, we do not expect their descriptors to match exactly, but rather correspond loosely. In this context,

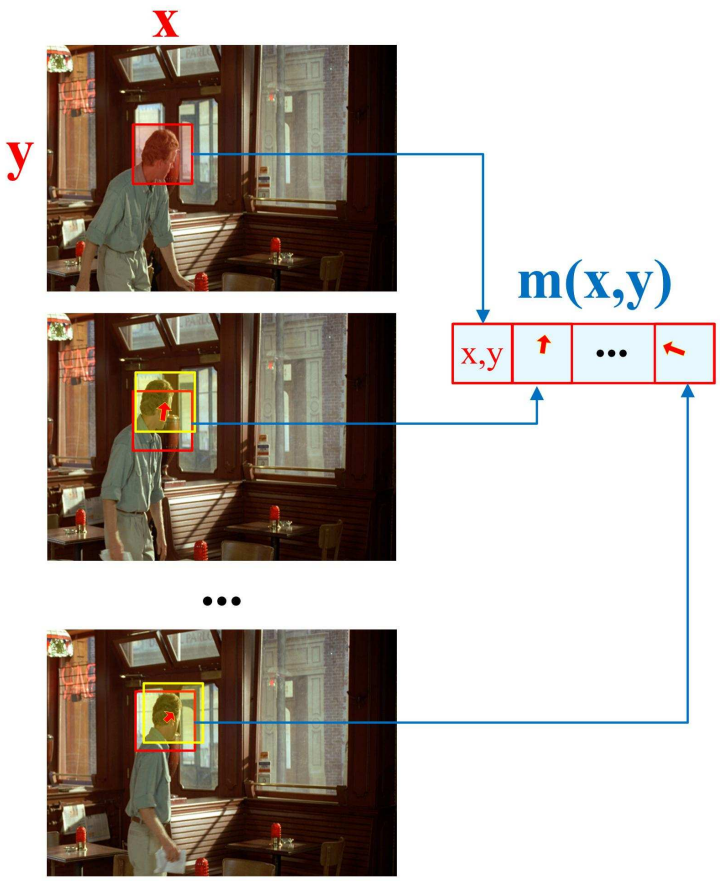

Fig. 2 Building a motion patch.

a statistical measure of the dissimilarity of the different sets of descriptors is adapted. In particular, entropic measures have proved relevant for image indexation [2,8]. We use the Kullback-Leibler (KL) divergence (noted D_{kl}) to evaluate the dissimilarity between the probability density functions (pdf) of each set of descriptors of the query and reference GoP (reminder: $D_{kl}(p_1||p_2) = \int p_1 \log(p_1/p_2)$). Noting $p_j(G)$ the pdf of the set $\mathbf{W}_j^G$ of spatial descriptors at scale j of GoP G and $p_m(G)$ the pdf of its set M^G of temporal descriptors, we thus consider the following dissimilarity measure:

$$D(G_q, G_r) = \alpha_1 \overbrace{D_s(G_q, G_r)}^{\text{spatial term}} + \alpha_2 \overbrace{D_t(G_q, G_r)}^{\text{temporal term}} \tag{4}$$

with

$$D_t(G_q, G_r) = D_{kl}\big(p_m(G_q)||p_m(G_r)\big) \tag{5}$$

$$D_s(G_q, G_r) = \sum_{j \geq j_0} D_{kl}\big(p_j(G_q)||p_j(G_r)\big). \tag{6}$$

The parameters α_1 and α_2 allow us to modulate the influence of the spatial versus the temporal term ($\alpha_1,\ \alpha_2 \geq 0$). j_0 is the coarsest scale of the decomposition (low-pass subband).

The temporal part of the dissimilarity measure ($D_t(G_q, G_r)$) compares the pdfs of the motion patches. Note that since those contain motion vectors plus their location (x,y), this term does not only indicate whether the sets of motions vectors are similar through time but it also takes into account whether they are organized similarly through space (indicating roughly whether similar shapes move the same way). A single spatial term in

$D_s(G_q, G_r)$ at scale j indicates whether local structures of spatial scale j are similar in the key frames of the two compared GoPs. Thus their sum ($D_s(G_q, G_r)$) indicates whether similar objects are present.

3.2 Computing the KL divergence

The dimension of our descriptors (both spatial and temporal) is high (from 16 to 27). Estimating the pdf and a fortiori the KL divergence in these large dimensions is not easy for at least two reasons: 1) in high dimensions, there is a lack of samples to accurately recover the pdf and 2) there is no multidimensional parametric models of the pdf that would both reflect the dependencies in our patches and allow for an analytic expression of the KL divergence in function of the model parameters. To alleviate both problems, we estimate the KL divergences in Eq. (4) directly, without estimating first the pdfs and without imposing a model on the pdf (this is a non-parametric model) by using a k-th Nearest Neighbor (kNN) approach.

This amounts to combining the Ahmad-Lin approximation of the entropies necessary to compute the divergences with "balloon estimates" of the pdfs using the kNN approach [11]. This is a dual approach to the fixed size kernel methods and was firstly proposed in [4]: the kernel bandwidth adapts to the local sample density by letting the kernel contain exactly k neighbors of x among a given sample set, so that the estimated pdf $\hat{p}$ from a sample set $\mathscr{W}$ reads:

$$\hat{p}(x) = \sum_{\mathbf{w}\in\mathscr{W}} \frac{1}{v_d\ \rho_{k,\mathscr{W}}^d(x)}\ \delta\left[||x-\mathbf{w}|| < \rho_{k,\mathscr{W}}(x)\right] \tag{7}$$

with v_d the volume of the unit sphere in $\mathbb{R}^d$ and $\rho_{k,\mathscr{W}}(x)$ the distance of x to its k-th nearest neighbor in $\mathscr{W}$. Plugging Eq.(7) in the Ahmad-Lin (cross-)entropy estimators and correcting for the bias, we obtain the following estimators of the KL divergence between two sets of d-dimensional points $\mathscr{W}_1$ and $\mathscr{W}_2$ of underlying pdf p_1 and p_2 (and containing N_1 and N_2 points) [1]:

$$\begin{aligned} D_{kl}(p_1||p_2) = &\log\left[\tfrac{N_2}{N_1-1}\right] + \tfrac{d}{N_1}\textstyle\sum_{n=1}^{N_1}\log[\rho_{k,\mathscr{W}^2}(\mathbf{w}_n^1)] \\ &- \tfrac{d}{N_1}\textstyle\sum_{n=1}^{N_1}\log[\rho_{k,\mathscr{W}^1}(\mathbf{w}_n^1)]. \end{aligned} \tag{8}$$

Note that this estimator is robust to the choice of k. For more details on the derivation of this estimators, we refer the reader to [1,7,8] and the references therein.

3.3 Scalability of the method

In this paper, we consider the problem of scalability of the measure in the following sense. We assume that the videos are available to the user through a heterogeneous network. Different persons thus may download the same videos under different format, e.g. using their PDA or their personal computer. More precisely, we assume that different users may download the same video with different levels of resolution; this is done by decoding more or less scales in the SVC stream for example. We consider that we know the minimal encoded resolution j_0.

We expect our dissimilarity measure to be robust to spatial resolution changes, assuming that the time resolution remains the same. This means that users having different versions of the same video should obtain similar answers to the same query submitted to the server. Indeed, the motion part of the dissimilarity is computed on large blocks corresponding to the lowest scale j_0 which is assumed to be the same for all users. On the opposite, the spatial part of the dissimilarity measure involves all scales j, some of which are not always accessible to the user. The sum in the spatial part of the dissimilarity is truncated to the scale available to the user. This truncation yields coherent result (see [7]) when comparing images. Theoretically, we thus obtain a spatially scalable measure. The experiments presented in Section 4 confirm that the proposed dissimilarity is robust to changes of resolution and hence is spatially scalable. Note that the temporal scalability is not taken care of here (i.e. when the temporal resolution changes with the format).

4 Experiments

In this section we provide some initial results of our GoP similarity measure. These experiments were performed on two video sequences from the ICOS-HD project database. After a brief description of the database, we present results of retrieval based on either spatial frame descriptors or on temporal/motion descriptors or on both sets of descriptors.

4.1 ICOS-HD video database

The ICOS-HD project[1] provides a large database of both original and re-edited video sequences. We used two of these sequences to test our similarity measure: *"Man in Restaurant"* (*S1*) and *"Street with trees and bicycle"* (*S2*)[2]. (Thumbnails of the two sequences are shown in Figure 3.)

S1 *S2*

Fig. 3 Thumbnails of the video sequences *S1* "Man in Restaurant"and *S2* "Street with trees and bicycle".

Each original sequence contains 72 Full HD frames (1920×1080 pixels) and has been manually split up into two clips, such that the boundary between the two clips roughly corresponds to a relevant motion transition, e.g. direction change of movement of an object or person. In addition, some common geometric and radiometric deformations were applied to the original HD video sequences, thus obtaining different versions of each video clip. In this paper we consider only two of these transformations: either a scaling to lower frame definition; or a quality degradation by high JPEG2000 compression. Each transformation was applied with two levels, as a result we used five different versions of each video sequence:

- original Full HD (1920×1080 pixels), referenced as 1920 in the figures;
- two rescaled versions (960×540 and 480×270 pixels), referenced as 960 and 540;
- two JPEG2000 coded versions (low and very low quality) referenced as jpeg2k 1 and jpeg2k2.

Each sequence being divided in two clips *C1* and *C2*, our test set contained exactly ten clips for each sequence.

As explained in Section 2, we used GoPs of 8 consecutive frames as basic units of video information to extract spatial and temporal descriptors for each clip. The spatial *SMP* descriptors were extracted from the first frame of each GoP using 4 resolution levels of the Laplacian pyramid as well as the low-frequency residual. The temporal descriptors were extracted using a diamond-search block matching algorithm to estimate inter-frame motion vectors on 16×16 blocks (corresponding to the lowest spatial resolution).

4.2 Spatial dissimilarity

In this paper we consider the task of retrieving the most similar GoPs to a query GoP. (Note that GoP retrieval can be easily generalized to retrieve even longer videos pieces, i.e. collections of consecutive frames, such as

[1] ICOS-HD (Scalable Joint Indexing and Compression for High-Definition Video Content) is a research project funded by ANR (French Research Agency).

[2] Original HD sequences ©Warner Bros issued from the Dolby 4-4-4 Film Content Kit One.

clips of multiple GoPs.) When performing this task, all transformed versions of the query GoP itself are expected to be ranked first by the dissimilarity measure defined above. For a query GoP G_q and a reference GoP G_r, the dissimilarity measure D defined in Eq. (4) is a combination of a spatial term D_s taking into account only spatial features and a temporal term D_t defined over temporal features. While spatial descriptors are essentially useful for comparing statistical scene information of two video pieces, motion descriptors are expected to highlight similarities based on dynamical patterns like the movement of objects or persons in a scene. The weighting factors α_1 and α_2 in Eq. (4) are used to privilege either term when performing a query.

Firstly we considered the case of $\alpha_1 = 1,\ \alpha_2 = 0$, i.e. only spatial descriptors were used to retrieve similar GoPs. In these experiments, the *SMP* descriptors proved to be crucial for distinguishing GoPs of the same video sequence as the query from those belonging to different video sequences. The results obtained are shown in Figure 4, where the dissimilarity of GoPs from both sequences is shown with respect to a query GoP taken from *S1*. (Namely the query is always the first GoP of the clip *C1* of sequence *S1*, in the 960 version. Each blue star in this figure is the dissimilarity to a particular reference GoP, which is identified by the sequence indicated in the middle of the figure, by the version of the sequence and the clip indicated on the x-label and finally by its occurrence in the clip, the 9 GoPs of a particular clip being ordered chronologically).

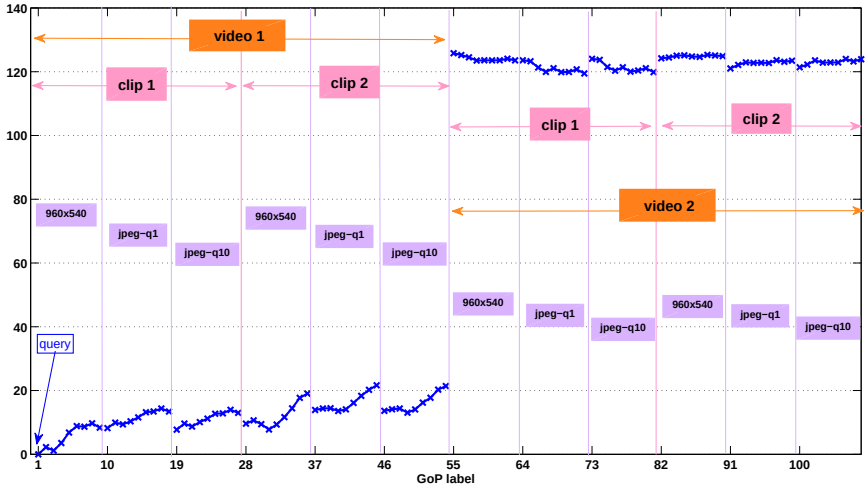

Fig. 4 GoP retrieval based on *SMP*. The query is GoP 1 from C1 of version 960 of *S1*.

Even when frame transformations are applied - either rescaling and very lossy compression - all GoPs originating from the same video sequence (*S1*) have small distances to the query, whereas all GoPs of sequence *S2* are far more dissimilar to the query. These results confirm that *SMP* descriptors are relevant for retrieving video scenes that share overall visual similarity with a query scene, and show in particular that the spatial part of the measure is robust to scaling and very lossy compression of a particular sequence (spatial scalability).

4.3 Temporal dissimilarity

We also tested the dissimilarity measure of Eq. (4) for $\alpha_1 = 0,\ \alpha_2 = 1$, i.e. when using only motion descriptors. Since the two clips of each sequence in our database differentiate from each other mainly for motion information, this measure is expected to discriminate GoPs of different clips of the same video sequence. This is confirmed by the experimental results shown in Figure 5, which show the motion dissimilarity from the query GoP (first GoP of the first clip of the 960 version of sequence *S1*) to all GoPs of the two clips of sequence *S1* in all versions (same labeling of the reference GoPs as for Fig. 4). The GoPs originating from clip *C1* (the same as the query) have far smaller dissimilarity values than those originating from clip *C2*, thus enabling the detection of a significant

Table 1 Mean and variance of the spatial and temporal dissmilarities

	Spatial term (across scenes)	Spatial term (within a scene)	Temporal term
Mean	122.8	12.1	3.7
Standard deviation	1.7	4.7	2.5

motion transition between the two clips. Note that the first two GoPs of clip *C2* are still not significantly dissimilar with respect to the previous ones, thus suggesting that such a manually detected transition is not abrupt. Indeed, the first clip corresponds to a continuous movement of the person from the scene center to the right side, whereas an inversion of movement direction (from right to left) occurs after the first few frames of the second clip. As previously, we note that the temporal part of the measure is robust to scaling and lossy compression (spatial scalability).

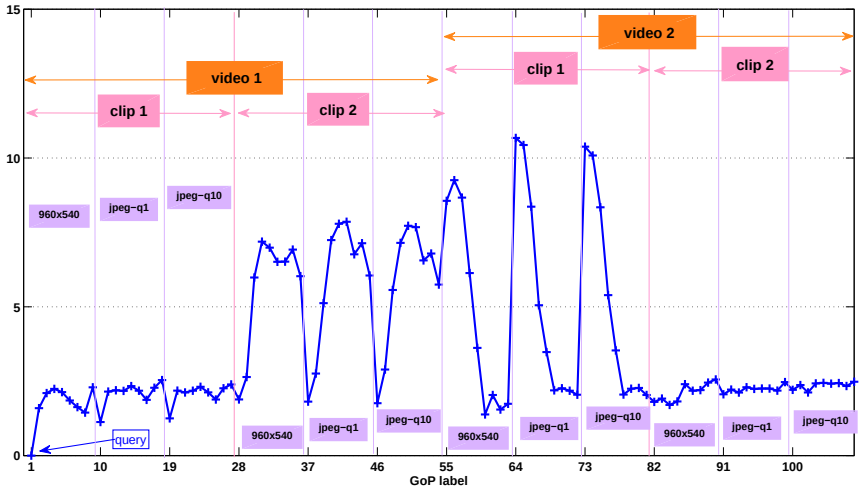

Fig. 5 GoP retrieval based on motion descriptors. The query is GoP 1 from C1 of version 960 of *S1*.

4.4 Spatio-temporal dissimilarity

In this section, we combine the spatial and temporal part of the dissimilarity measure to obtain a global dissimilarity. Considering that the spatial term of the dissimilarity is able to differentiate video scenes and the temporal term allows to characterize different motions within a single sequence, we expect that the combination of the two will enable to globally compare two clips whether there are or not from the same sequence.

The typical ranges and variances of the spatial and temporal similarity are quite different (see Table 1). As seen from the previous experiments, the spatial term is not discriminative within a scene but shows a clear discontinuity marking the difference between scenes, while the temporal term differentiates GoPs within a video. We thus rescale the temporal term to ensure that on average it modulates the spatial term within a scene without breaking the discontinuity across scenes. To do so, we set $\alpha_1 = 1,\ \alpha_2 = 10$. The results displayed in Fig. 6 indeed show that the two clips within a sequence are discriminated independently of which degradation is applied to the reference GoPs.

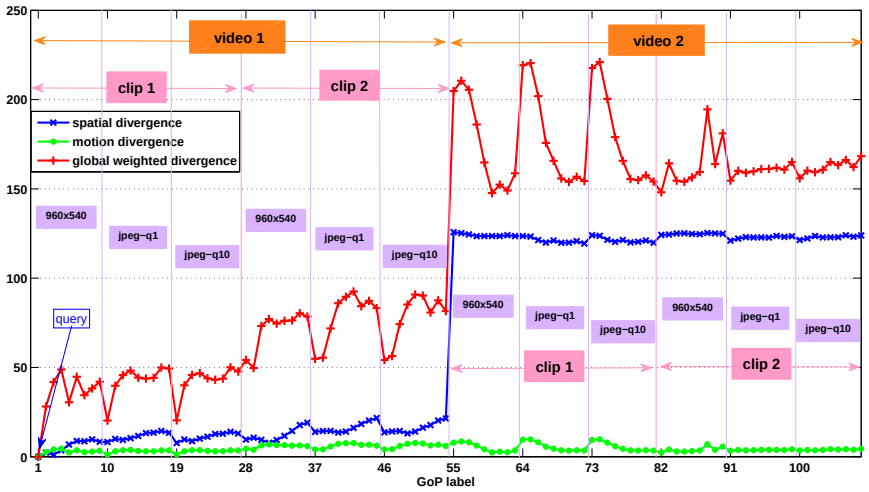

Fig. 6 GoP retrieval combining spatial (weight $\alpha_1 = 1$) and temporal (weight $\alpha_2 = 10$) dissimilarities. The query is GoP 1 from C1 of version 960 of *S1*. The reference GoP on the x-axis are ordered as in Fig. 4

5 Conclusion

In this paper, we have proposed both spatial and motion descriptors and a dissimilarity measure to compare video sequences. The basic unit to compare videos is the GoP (circa 8 frames). The spatial descriptors called *sparse multiscale patches* capture the visual information of a reference frame of the GoP in a multiscale fashion. The motion descriptors called *GoP motion patches* capture the motion in a GoP at the block level. Both kind of descriptors rely on the concept of patches i.e. groups of neighboring elements whose coherence is exploited in a statistical dissimilarity measure. To compare two GoPs, we propose a statistical measure that combines a spatial term and a temporal term. It is a sum of Kullback-Leibler divergences between pdfs of sets of spatial and temporal patches, that is estimated in a non-parametric setting via the k-th nearest neighbor framework.

The motion and a spatial terms of the dissimilarity measure were studied independently and jointly. The test set contained rescaled and compressed versions of two videos sequences divided into two clips that are characterized by different motion. The results obtained using either only spatial descriptors or only motion descriptors show that both terms are robust to these transformations. This indicates that the proposed measure contains the spatial scalability properties required to be coherent when used with the different data formats available on heterogeneous networks. The spatial term discriminates different video scenes while the temporal term discriminates different motion within a scene. The experiments using the full dissimilarity also show how weightings of the spatial and temporal parts of the dissimilarity measure allow to discriminate simultaneously different sequences and different clips within a sequence and confirms the spatial scalability of the method. These experiments suggest that, depending on the particular video retrieval task, a combination of both dissimilarity terms in Eq. (4) is relevant to detect similar video samples in a database containing both original and degraded versions of different video clips. Different search criteria may be targeted by adjusting the weights α_1, α_2, e.g. from searching similar movements of objects in a scene independently of the background to searching visually similar scenes ignoring the movement of objects or persons in the scene.

Acknowledgements The authors would like to acknowledge the contribution of W. Belhajali in the experiments conducted here.

References

1. 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.

2. Alfred O. Hero, Bing Ma, Olivier Michel, and John Gorman. Alpha-divergence for classification, indexing and retrieval. Technical Report CSPL-328, University of Michigan, 2001.
3. I. Laptev and P. Pérez. Retrieving actions in movies. In *Proc. Int. Conf. Comp. Vis.(ICCV'07)*, pages 1–8, Rio de Janeiro, Brazil, October 2007.
4. D. Loftsgaarden and C. Quesenberry. A nonparametric estimate of a multivariate density function. *AMS*, 36:1049–1051, 1965.
5. V. Mezaris, I. Kompatsiaris, and M. G. Strintzis. Object-based mpeg-2 video indexing and retrieval in a collaborative environment. *Multimed. Tools Appl.*, 30:255–272, 2006.
6. C. Morand, J. Benois-Pineau, J.-Ph. Domenger, and B. Mansencal. Object-based indexing of compressed video content: from sd to hd video. In *IEEE VMDL/ICIAP*, Modena, Italy, September 2007.
7. P. Piro, S. Anthoine, E. Debreuve, and M. Barlaud. Image retrieval via kullback-leibler divergence of patches of multiscale coefficients in the knn framework. In *CBMI*, London, UK, June 2008.
8. P. Piro, S. Anthoine, E. Debreuve, and M. Barlaud. Sparse multiscale patches for image processing. In *ETVC*, volume 5416/2009 of *LNCS*. Springer, 2009.
9. F. Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Segmenting, modeling, and matching video clips containing multiple moving objects. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 29(3):477–491, mar 2007.
10. H. Schwarz, D. Marpe, and T. Wiegand. Overview of the scalable video coding extension of the h.264/avc standard. *Circuits and Systems for Video Technology, IEEE Trans. on*, 17(9):1103–1120, Sept. 2007.
11. David W. Terrell, George R.and Scott. Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236–1265, 1992.
12. Y. Zhai, J. Liu, X. Cao, A. Basharat, A. Hakeem, S. Ali, and M. Shah. Video understanding and content-based retrieval. In *TRECVID05*, November 2005.
13. D. Zong and S.F. Chang. An integrated approach for content-based video object segmentation and retrieval. *IEEE Trans. On Circuits and Systems for Video Technologies*, 9:1259–1268, 1999.