Scalable spatio-temporal video indexing using sparse multiscale patches

Paolo Piro, Sandrine Anthoine, Eric Debreuve, Michel Barlaud

▶ **To cite this version:**

Paolo Piro, Sandrine Anthoine, Eric Debreuve, Michel Barlaud. Scalable spatio-temporal video indexing using sparse multiscale patches. CBMI '09, Jun 2009, Chania, Greece. pp.95-100, 10.1109/CBMI.2009.48 . hal-00417411

HAL Id: hal-00417411
https://hal.science/hal-00417411

Submitted on 15 Sep 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Published in proceedings of CBMI 2009, 03-05/06/2009, Chania, Greece.

Scalable spatio-temporal video indexing using sparse multiscale patches

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 address the problem of scalable video indexing. We propose a new framework combining sparse spatial multiscale patches and Group of Pictures (GoP) motion patches. The distributions of these sets of patches are compared via the Kullback-Leibler divergence estimated in a non-parametric framework using a k-th Nearest Neighbor (kNN) estimator. We evaluated this similarity measure on selected videos from the ICOS-HD ANR project, probing in particular its robustness to resampling and compression and thus showing its scalability on heterogeneous networks.

1 Introduction

In the last decades, the number of video databases available through different heterogeneous networks has grown rapidly, together with the need for efficient tools to order, explore and use these databases of videos with a variety of size and formats. In this paper, we tackle the problem of video indexing which consists in finding a suitable description of the video content for effective search in databases. The search is content-based meaning that no prior manual annotation has taken place on the video database. 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 [9]. 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. Other methods explicitly take into account the motion and visual information in the video. Amongst these are object based video indexing methods [10, 4, 7] that rely on a segmentation of the semantic objects in the video. The object is usually segmented spatially and its motion is followed through the video. An approach that relies on segmenting video sequences in spatio-temporal "volumes" has been recently proposed [1]. This approach aims to extract features from relevant spatio-temporal regions of a video scene and match them to find similar videos.

Recognizing spatio-temporal similarities between video sequences is a difficult task, 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 task on video databases (e.g. find videos with similar motion or videos containing a similar object) and 2) to provide coherent answers with the various data formats that are available to the user through a heterogeneous network (i.e. give similar answer whether the user is sending its query from a PDA or his desktop computer). To do so, we define two kinds of descriptors, 1) global visual descriptors - also called spatial descriptors - that capture the visual content of a scene and 2) temporal or motion descriptors that capture the trajectories of objects in the videos. Both kinds of descriptors are patch descriptors that exploit respectively the spatial and 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.

Section 2 specifies the descriptors we propose (both spatial and temporal). The dissimilarity measure is defined in Section 3, and its practical implementation using the k-th nearest neighbors approach is detailed. 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)* that characterize the visual features of still images (see [5, 6]). 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.

*This work is supported by the French ANR grant "ICOS-HD".

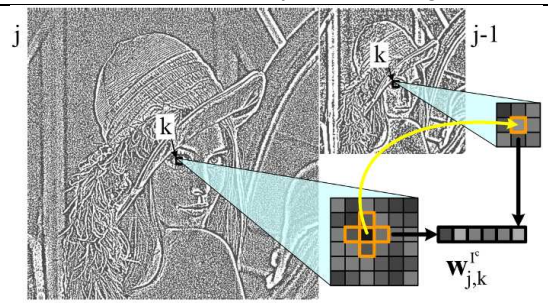

Figure 1. Building a patch of multiscale coefficients, for a single color channel image.

2.1 Spatial descriptors: sparse multiscale patches (*SMP*)

We rapidly present here our spatial descriptors described in details in [5, 6]. Each *SMP* of an image captures the local structure of a given scale at a specific location since its contains multiscale coefficients of all color channels of the image that are neighbors across scale and location. More precisely, noting $w^{I^c}_{j,k}$ the coefficient of channel c of image I at scale j and 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):

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

Interchannel patches $\mathbf{W}^{I}_{j,k}$ 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 $\mathbf{w}^{I^V}_{j,k}$:

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

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. 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. (From now on, $\mathbf{W}^{I}_{j,k}$ will denote 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, 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 $\{\mathbf{W}^{I}_{j,k}\}_{j\geq j_0, k\in\mathbb{Z}}$ that describes well the content of the image. 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.

2.2 Temporal descriptors: GoP motion patches (*GoP-MP*)

To describe accurately the motion of objects in a video, we also use the concept of patches. Here, the patches are understood as group 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. To do so, we compute the apparent motion of each particular block (x,y) through a short period of time, typically a GoP (of around 8 to 10 pictures). This way, each block (x,y) is bound to belong to a single object that we follow through the GoP. 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):

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

where $\mathbf{u}_{n-1,n}$(x,y) is the apparent motion of point (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). 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 location (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)). 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 measure 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

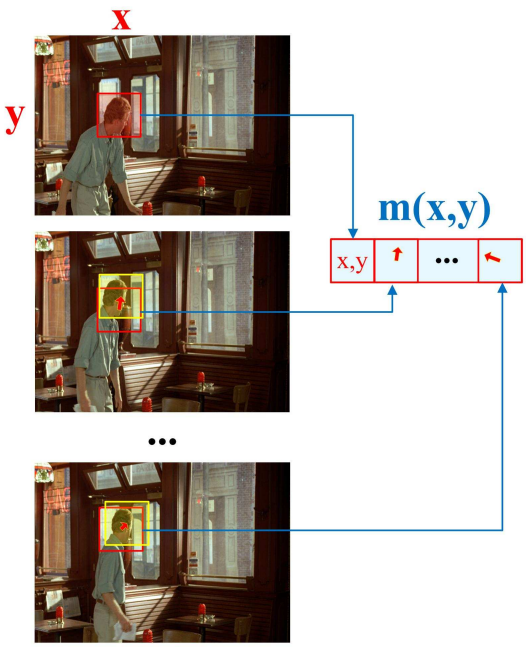

Figure 2. Building a motion patch.

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, a statistical measure of the dissimilarity of the different sets of descriptors is adapted. 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 $\mathbf{W}^G_j$ 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:

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

with

$$\begin{aligned} D_t(G_q, G_r) &= D_{kl}\big(p_m(G_q)||p_m(G_r)\big) \\ D_s(G_q, G_r) &= \sum_{j \geq j_0} D_{kl}\big(p_j(G_q)||p_j(G_r)\big). \end{aligned}$$

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).

3.2 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 that minimal encoded resolution j_0.

We expect our dissimilarity measure to be robust to resolution changes, meaning 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 which is the same for all users, while the sum in the spatial part of the dissimilarity can be truncated to the scale available to the user (we showed in [5] that these truncations yield coherent results). The experiments presented in Section 4 also show that the proposed dissimilarity is robust to changes of resolution and hence is scalable.

3.3 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 [8]. This is a dual approach to the fixed size kernel methods and was firstly proposed in [3]: 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 $\mathcal{W}$ reads:

$$\hat{p}(x) = \sum_{\mathbf{w}\in\mathcal{W}} \frac{1}{v_d\ \rho_{k,\mathcal{W}}^d(x)}\ \delta\big[||x-\mathbf{w}|| < \rho_{k,\mathcal{W}}(x)\big] \quad (5)$$

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 $\mathcal{W}$. Plugging Eq.(5) 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 $\mathcal{W}_1$ and $\mathcal{W}_2$ of underlying pdf p_1 and p_2 (and containing N_1 and N_2 points) [2]:

$$D_{kl}(p_1||p_2) = \log\Big[\tfrac{N_2}{N_1-1}\Big] + \tfrac{d}{N_1}\textstyle\sum_{n=1}^{N_1}\log[\rho_{k,\mathcal{W}^2}(\mathbf{w}_n^1)] - \tfrac{d}{N_1}\textstyle\sum_{n=1}^{N_1}\log[\rho_{k,\mathcal{W}^1}(\mathbf{w}_n^1)]. \quad (6)$$

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 [5, 6] and the references therein.

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.

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.)

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

S1 *S2*

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

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. 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. Temporal descriptors were extracted using a diamond-search block matching algorithm to estimate inter-frame motion vectors on 16×16 blocks.

4.2 GoP similarity using spatial *SMP* descriptors

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 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

[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.

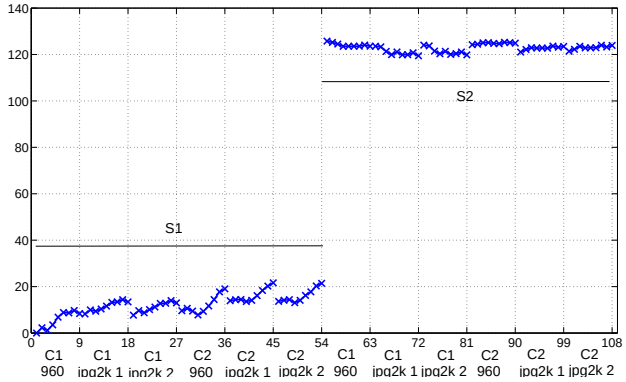

Figure 4. GoP retrieval based on *SMP*. The query is GoP 1 from C1 of version 960 of *S1*.

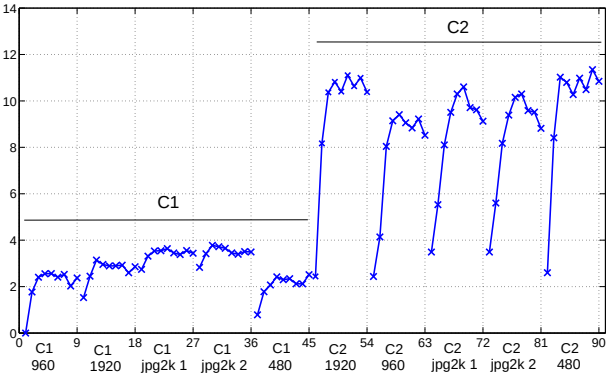

Figure 5. GoP retrieval based on motion descriptors. The query is GoP 1 from C1 of version 960 of *S1*.

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 occurence in the clip, the 9 GoPs of a particular clip being ordered chronologically).

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 suitable 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.

4.3 Similarity using motion descriptors

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 labelling of the GoPs as for Fig. 4). 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 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). Namely, 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.

4.4 Combining spatial and temporal similarities

We also tested a combination of the spatial and temporal terms as a global dissimilarity measure.

We tested two sets of combinations. For the first one, we set $\alpha_1 = 1,\ \alpha_2 = 1$. Fig. 6 (top) shows that, in this case, the motion terms mildly modulate the spatial terms and as a result the global dissimilarity is able to discriminate Clips within and between sequence. This equal weighting of the terms stresses the difference between GoPs from different sequences: this corresponds to the fact that the scale of the spatial term is intrinsically larger than the motion terms. To take full advantage of the motion term, its weighting in increased: we set $\alpha_1 = 1,\ \alpha_2 = 3$. The results displayed in Fig. 6 (bottom) for this case indeed show that the two Clips within a sequence are discriminated independently of which

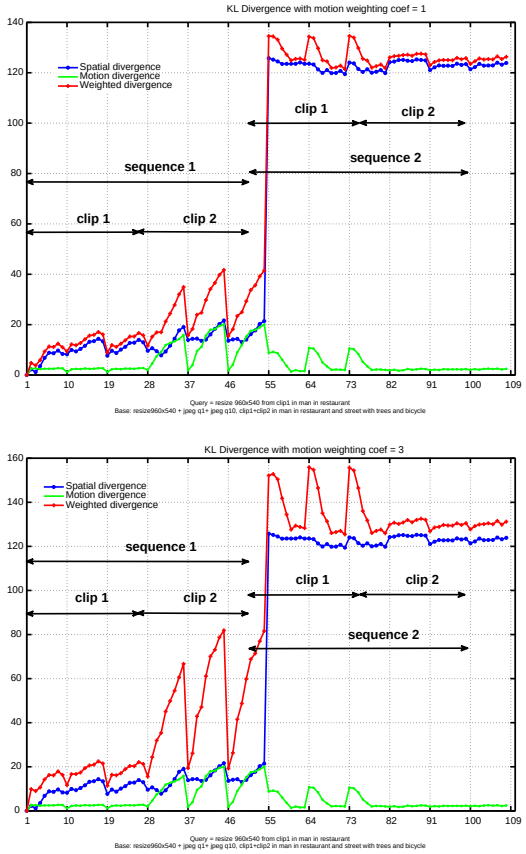

Figure 6. GoP retrieval combining spatial (weight α_1) and temporal (weight α_2) dissimilarities. The query is GoP 1 from C1 of version 960 of *S1*. Top: equal weights $\alpha_1 = \alpha_2 = 1$. Bottom: $\alpha_1 = 1,\ \alpha_2 = 3$. The reference GoP on the x-axis are ordered as in Fig. 4

degradation is applied to the reference GoPs.

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 in a frame. The motion descriptors called *GoP motion patches* capture the object trajectories through a GoP. 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. This measure is a sum of Kullback-Leibler divergences between pdfs of sets of patches, that is estimated in a non-parametric setting via the k-th nearest neighbor framework.

The dissimilarity contains a motion and a spatial term that were studied independently on a test set of rescaled and compressed versions of two videos sequences divided into two clips. 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 scalability properties required to be coherent when used with the different data formats available on heterogeneous networks. The experiments also suggest that, depending on the particular video retrieval task, a combination of both dissimilarity terms in Eq. (4) is suitable 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 $\alpha_1,\ \alpha_2$, 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.

References

[1] A. Basharat, Y. Zhai, and M. Shah. Content based video matching using spatiotemporal volumes. *Computer Vision and Image Understanding*, 110(3):360–377, 2008.

[2] 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.

[3] D. Loftsgaarden and C. Quesenberry. A nonparametric estimate of a multivariate density function. *AMS*, 36:1049–1051, 1965.

[4] C. Morand, J. Benois-Pineau, J.-P. Domenger, and B. Mansencal. Object-based indexing of compressed video content: from sd to hd video. In *IEEE VMDL/ICIAP*, Modena, Italy, September 2007.

[5] 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.

[6] P. Piro, S. Anthoine, E. Debreuve, and M. Barlaud. Sparse multiscale patches for image processing. In *ETVC*, volume 5416 of *LNCS*. Springer, 2009.

[7] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. Segmenting, modeling, and matching video clips containing multiple moving objects. *IEEE PAMI*, 29:477–491, 2007.

[8] D. W. Terrell, George R.and Scott. Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236–1265, 1992.

[9] 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.

[10] D. Zong and S. 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.