Efficient Spatio-Temporal Edge Descriptor

Claudiu Tanase, Bernard Merialdo

Efficient Spatio-temporal Edge Descriptor

Claudiu Tănase[1] and Bernard Mérialdo[1]

EURECOM, 2229 Route des Crêtes, Sophia-Antipolis, France

Abstract. Concept-based video retrieval is a developing area of current multimedia content analysis research. The use of spatio-temporal descriptors in content-based video retrieval has always seemed like a promising way to bridge the semantic gap problem in ways that typical visual retrieval methods cannot. In this paper we propose a spatio-temporal descriptor called ST-MP7EH which can address some of the challenges encountered in practical systems and we present our experimental results in support of our participation at TRECVid 2011 Semantic Indexing. This descriptor combines the MPEG-7 Edge Histogram descriptor with motion information and is designed to be computationally efficient, scalable and highly parallel. We show that our descriptor performs well in SVM classification compared to a baseline spatio-temporal descriptor, which is inspired by some of the state-of-the-art systems that make the top lists of TRECVid. We highlight the importance of the temporal component by comparing to the initial edge histogram descriptor and the potential of feature fusion with other classifiers.

Keywords: spatio-temporal, descriptor, content-based video retrieval, high-level feature extraction, classification, concept, edge histogram

1 Introduction

High-level feature extraction is a fundamental topic in multimedia research. Also known as semantic concept detection, its goal is to determine the presence or absence of semantic concepts in multimedia content. We investigate the presence of such concepts in video shots by using *concept classifiers*, which measure the relevance of a concept within a video shot. In the literature, much of the work in this domain is tested against the TRECVid benchmark, which provides large video databases, manual concept annotation that can be used for training and standardized evaluation measures, such as the widely used Mean Average Precision [1]. The task of each participant is to build a system that automatically identifies the video shots where a particular concept is shown (e.g. there is an occurence of concept 'dog' in shot 10 of video 244 in the set), and then rank them by relevance. The database is divided in the training set, which is the base for the experiments, and a test set, on which the system performances will be evaluated. Each participant must provide a list of 2000 shot IDs ranked by decreasing probability of detecting the particular concept. TRECVid organizers provide a shot decomposition, as well as a central keyframe for each shot of the video.

Highly performing [1] systems in TRECVid rely almost exclusively on image descriptors, computed over central keyframes in shot, so they are basically classifying and retrieving images. However, some [2] sample several keyframes in one shot, and others use local features computed around spatio-temporal interest points (STIP [3]) within the video. In spite of this, in the recent editions of the TRECVid Semantic Indexing task (SIN), progress seems to have slowed down because of several flaws in learning methods or dataset/annotation problems [4]. As each year more and more dynamic and motion-relevant concepts are added, the use of one or few keyframes per shot in concept detection is beginning to show its limitations.

Spatio-temporal descriptors have been used for various video detection tasks, most notably in human action recognition. These descriptors show very good discrimative features and are reliable in general, but have steeper computing requirements and sometimes need pre-processed video data [5,6]. For these reasons, their adoption in real systems at TRECVid has been slow and with mediocre practical results [1]. The most likely cause is the comparatively high computational cost that comes with descriptors on *xyt* space, but also because of dataset problems, such as high intra-concept variability and very few positive instances of concepts.

We propose a spatio-temporal descriptor that can work around these problems. Our ST-MP7EH descriptor is based on the Edge Histogram image descriptor, part of the MPEG-7 standard [7], and is basically analyzing the temporal evolution of edges in video. Our descriptor works by computing an edge histogram in each frame, and then calculating two simple statistic parameters on the distribution in time of each "bin" in the histogram. By subsampling frames at a reasonably low rate we can decrease computation time, which is essential given our large video datasets (TRECVid2010 has 200 hours of video for training and testing). This does little to impact the quality of the descriptor since our temporal statistics (moments) are theoretically invariant to this operation.

This paper is structured as follows: in section 2 we present some existing interesting approaches that rely on spatio-temporal detection, specially using edges. We present the MPEG-7 Edge Histogram, which is the starting point for our work, in section 3. In section 4 we present the method for our descriptor, while also motivating our design decisions. In section 5 we present the details of our testing and comparisons, we show how our spatio-temporal SIFT baseline is constructed and in section 6 we show results computed on actual TRECVid data. We conclude our paper with our comments on the descriptor's performance and future use in section 7.

2 Previous Work

In this section we describe known spatio-temporal video retrieval techniques that use features based on edges and have been successfully used in TRECVid or are part of the state of the art in concept video classification or action recognition. TRECVid systems seem to tend toward increasing the number of visual features and incorporating sophisticated fusion strategies while relying less on motion or edge information [1]. The most successful systems do incorporate spatio-temporal information by sampling multiple keyframes, however their number is extremely limited (MediaMill uses up to 6 additional I-frames distributed around the middle key frame of each shot) [2].

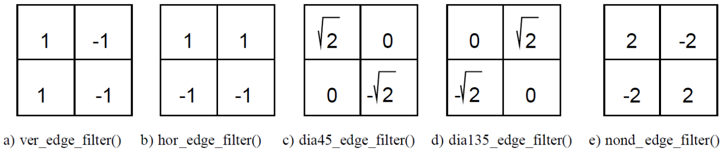
However, several TRECVid participants, mostly in the new Multimedia Event Detection (MED), do use edge features. In SIN (high level feature extraction), the MPEG-7 Edge Histogram has been used only in systems that work with the middle keyframe of the shot [8–10], thus without any spatio-temporal or motion information. On the other hand [11] compute edge histograms on a local level and use the BoSW (Bag of Spatiotemporal Words) strategy to track features in the space-time volume. An interesting approach is the TGC (temporal gradient correlogram) by [12], which computes edge direction probabilities over 20 frames evenly distributed in the shot. Their work is similar in principle to ours, but the temporal aspect is represented by a mere concatenation of the 20 vectors resulting from the 20 sampled shots. In spite of some temporal information, this approach is highly dependent on the shot length and has a higher computational cost. The EOAC [13] (edge orientation autocorrelogram) is practically identical. Another example of edge-based descriptor is MEHI (motion edge history image) from [14], which evolves from previous MHI and MEI (heavily used in human action recognition), and is a suitable descriptor for human activity detection. However, its use in general concept-based retrieval is questionable because of camera motion, broader range of possible motions and inherent video quality problems that come with Internet archive videos. Moreover, the exhaustive manner of computation could prove impractical for the high-level feature task.

3 MPEG-7 Edge Histogram Descriptor

The MPEG-7 standard describes an Edge Histogram descriptor for images, which is meant to capture the spatial distribution of edges, as part of a general texture representation. As with all color and texture descriptors defined in the MPEG-7 standard [7], this descriptor is evaluated for its effectiveness in similarity retrieval [15], as well as extraction, storage and representation complexity. The distribution of edges is a good texture signature that is useful for image to image matching even when the underlying texture is not homogeneous.

The exact method of computation for the MPEG-7 Edge Histogram descriptor can be found in [7] and [15]. The general idea is that the image is divided into 4×4 sub-images, and the local edge histograms are computed for each of the sub-images. There are 5 possible edge orientations that are considered: vertical, horizontal, 45° diagonal, 135° diagonal and isotropic (no orientation detected). For each sub-image and for each image type an edge intensity bin is computed, amounting to a total of $16 \ images \times 5 \ edges = 80$ bins.

Each sub-image is further divided into sub-blocks, which are down-sampled into a 2×2 pixel image by intensity averaging, and the edge-detector operators are applied using the 5 filters in the image below. The image blocks whose edge strengths exceed a threshold are marked as ”edge blocks” and used in computing the histogram. These values are counted and normalized to $[0, 1]$ for each of the 80 bins. The value in each bin represents the ”strength” of the corresponding edge type in that image block. According to its authors [7], this image descriptor is effective for representing natural images for image-to-image retrieval. It is not suited for object-based image retrieval. Moreover, the computation is efficient [15], and has low dimensionality and storage needs.



a) ver_edge_filter() b) hor_edge_filter() c) dia45_edge_filter() d) dia135_edge_filter() e) nond_edge_filter()

Fig. 1. MPEG-7 directional filters [15]

4 ST-MP7EH Spatio-Temporal Descriptor

In the context of video retrieval, visual descriptors that are traditionally used in CBIR are sometimes used to describe frame sequences instead of images, following a more or less elaborate extension process. A good example of a properly built 3D descriptor is the 3D extension [16] of SIFT (evidently the highest performing visual feature in image search) or the extension [17] of HOG used initially in human action detection. However these cases are rare, as most systems tend to use keyframe-based approaches or compute 2D descriptors at salient points in the ST volume (detected by spatio-temporal interest points).

Our opinion is that temporal and spatio-temporal features have a huge potential in content-based video retrieval, and at the same time we state that keyframe approaches miss a great deal of information on two aspects: the feature we are looking for may not be present on the selected keyframe (but present on other frames in the shot), and the feature is easier to recognize by means of its dynamics throughout the shot rather than its instantaneous visual characteristics [18]. Naturally ST methods require far more processing power than keyframe approaches, so a compromise between performance and computational cost must always be made. For that reason, we decided on MPEG-7 Edge Histogram thanks to its low computational cost [15].

In the same idea as the seminal work of Nelson and Polana [19] we tried to create a global and general descriptor that can detect the evolution in time of visual texture. In our case, the texture is characterized by the predominance of an oriented edge on a region of the image. The ST descriptor is computed in a simple manner: for each frame t of the analyzed video, we compute the

(2D) MPEG-7 edge histogram descriptor, which gives an 80 value feature vector x_1^t to x_{80}^t. We compute this on every frame and put the data in an $N \times 80$ matrix, where N is the number of analyzed frames. We consider each column j of this matrix as a time series $x_j^1, x_j^2, ..., x_j^N$. This series represents the evolution in time of a certain feature of the image (e.g. the 19th element represents the strength of horizontal edges on the 3rd sub-image). The feature vector is made from the average a_j and standard deviation σ_j of each of these 80 series, namely $[a_1, a_2, ..., a_N, \sigma_1, \sigma_2, ..., \sigma_N]$ which gives it a fixed dimension of 160. The order of magnitude is conserved between the two descriptors by means of average and variance. The spatial information given by the edge distribution and the division into grids is inherited from the underlying edge histogram descriptor. Temporal information is present in the form of a shot-wide average edge histogram plus the temporal variance in each histogram bin. Formally, the mean represents the first moment of the discrete distribution, and the standard deviation is the square root of the second central moment (the variance). We use the standard deviation and not the variance in order to conserve the metric (to have the same measuring unit), which is essential to classification.

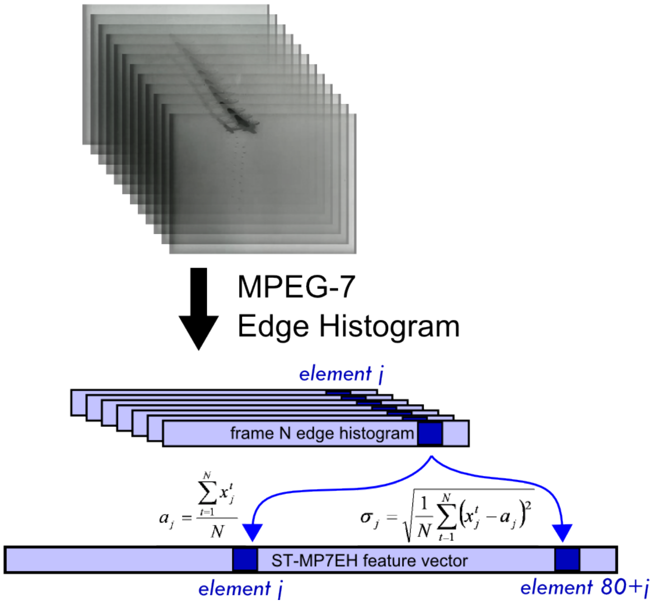

MPEG-7
Edge Histogram
element j
frame N edge histogram
$a_j = \frac{\sum_{t=1}^{N} x_j^t}{N}$
$\sigma_j = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(x_j^t - a_j\right)^2}$
ST-MP7EH feature vector
element j
element 80+j

Fig. 2. Overview of ST-MP7EH computation

ST-MP7EH has some interesting temporal properties. Firstly, it is a direct temporal extension for the MPEG-7 Edge Histogram: if we compute our descriptor on a sequence containing just one frame N=1, the resulting feature vector would hold the edge histogram for that frame, with all the extra variances equal to zero. We find this helpful in providing a comparison between the 2D descriptors extensively used in video retrieval and the corresponding 3D ST extensions. Secondly, this descriptor is robust to frame subsampling, as means and variances are statistically invariant to subsampling. Thirdly, this is a one-pass global descriptor, which is to say that concerning memory, it only touches each 3D point (in the XYT space) only once, and does so in a linear fashion. The consequence is that sliding-window implementations are simple and efficient (we can cut the ST volume anywhere on the T axis), that computation time is easy to estimate (directly proportional to the number of frames) and that its "global" property ensures that no region in the ST volume will be missed because of bad STIP detection [3], for instance.

Similar temporal strategies have been used in space-time analysis before, most notably in human gesture recognition. Darrell and Pentland propose the use of Dynamic Time Warping [20] for gesture recognition, and other authors have proposed frequency domain [19] (Fourier analysis) and wavelets [21] to detect repetitive patterns in walking motion. However, in the context of general motion of an object in a video sequence, possibly combined with noise and camera motion, periodicity would obviously not prove as robust. The computational overhead would also be significant by comparison to state-of-the-art video retrieval systems. One thing our approach shares in common with frequency domain representations is that both methods store the mean of the signal: ST-MP7EH computes it explicitly and Fourier analysis computes the mean as the first Fourier coefficient.

In order to minimize computation time we temporally sub-sampled the frames of the shot by a 1/5 ratio. We found the subsampling appropriate for two reasons: firstly, because the functions we use should be unaffected by the subsampling of the dataset, and secondly because we assume the continuity of the MPEG-7 edge strengths in time (as they are calculated from a continuous shot). Intuitively this property should hold true for a sequence of frames forming a continuous shot: since the difference between any 2 consecutive shots is small, so should be the difference between 2 elements of the edge histogram for the corresponding edges.

5 Experimental Setup

We have tested our descriptors on our TRECVid 2011 testbed platform. Eurecom's system uses a fusion of several visual classifiers (SIFT, GIST, color moments, wavelet features), of which SIFT is the earliest and still the most powerful, just like in most TRECVid systems. ST-MP7EH has been designed to complement the image descriptors by providing spatio-temporal information which would be invisible to keyframe-based descriptors. A computational constraint is also imposed by the amount of video data that is exponentially increasing from one edition of TRECVid to the next and the available hardware. The scoring metric for TRECVid SIN is the MAP (Mean Average Precision). Average precision (AP) is a standard performance metric of a concept classifier. Given the classifier's output as relevance scores on a set of shots, we rank the shots in descending order of their score, and compute AP as the average of the precisions of this ranked list truncated at each of the relevant shots. The mean of APs, or mean average precision (MAP), is a metric of the average performance of multiple concept classifiers.

5.1 ST-MP7EH Evaluation

We test the performance of our ST-MP7EH descriptor on TRECVid data. The chosen training and test sets are two subsets of the annotated TRECVid2010 data available at the moment of writing. Our experiments were conducted using the 10 concepts from the TRECVid 2010 "light run" of the SIN task. The training set contains 59800 shots and the test set 59885 shots. The video data comes from approximately 8000 Internet Archive videos (50GB, 200 hours) with Creative Commons licenses in MPEG-4/H.264 with durations between 10 seconds and 3.5 minutes. We compute our descriptor on all available shots, which are segmented using an automated boundary detection mechanism. Given the fact that this is an "embarrassingly parallel" problem, splitting the workload into a manageable number of jobs is trivial. Computation time for a single shot depends on shot length and frame size, as well as hardware-dependent considerations. On average for 10% of the tv10.shorts.B corpus (137327 shots) it takes approx. 28.23 hours, which makes for an average 7.4011 seconds per shot. This can also be approximated as 1.23× playback time. We estimate memory usage as 5.747 kB per shot.

The next step is the SVM training. We label our data using the available annotation. At this point the number of training examples becomes concept-dependent as the annotation is not complete over the entire dataset. We use a modified version of the SVM software available from LibSVM [22] that uses the $\chi 2$ kernel. As with all other SVM training experiments, we use one single-class SVM per concept that should differentiate between positive and negative samples. Given the disproportionate nature of the positive and negative examples (positive/negative ratio is $< 1\%$ for every concept), the label obtained in testing will always be negative. We use the assigned soft-boundary probability P as an indication of how likely the test vector is to actually be a positive instance, and call this the "score" of the shot. We sort by this probability in order to obtain our ranked list on which we compute the AP for the 10 concepts. The average of APs over all concepts is the MAP.

5.2 Comparison with Spatio-Temporal Baseline Descriptors

We compare our descriptor with several baselines in terms of MAP. The first experiment is meant to compare direct retrieval quality, regardless of the nature of the descriptor. According to [1], the best individual visual descriptor in the current generation of video concept detection systems is still SIFT. We use a Bag of Words approach by clustering all SIFT features into 500 clusters (visual words) and constructing a histogram of visual word occurrences for each sample, based on the nearest visual word. Inspired by the work of MediaMill [2], we adopt a multi-keyframe approach, were we sample a large number of keyframes from the shot (1/5 regular frames), compute SIFT features, and finally create a single visual word occurrence histogram per keyframe. We classify using the same SVM method as for ST-MP7EH. At this point the 3 variants of our descriptor branch: we either consider the highest-scoring keyframe as the overall score for the corresponding shot (*mkfSIFT1*), average all scores to get the shot score (*mkfSIFT3*), or average the visual word histograms (*mkfSIFT2*) and finally compute the MAP. We consider these baseline descriptors as prototype spatio-temporal visual detectors for general concept classification. We motivate this by highlighting the fact that none of the descriptors that work well on human action [20, 21], surveillance, etc. have made their way into general-concept systems because of their weak generalizing power.

5.3 Spatio-Temporal Performance Gain

We highlight the temporal quality of ST-MP7EH by directly comparing retrieval performance to its "predecessor", the MPEG-7 Edge Histogram. For that we compute Edge Histograms on one relevant keyframe in each shot and use the resulting 80-value vector in SVM classification. Results clearly indicate the gain of using multiple keyframes per shot. This experiment has been carried out on a subset of the training set, containing half of the training samples.

5.4 ST-MP7EH - SIFT Late Fusion

Following the multiple descriptor fusion paradigm that seems to dominate current state-of-the-art systems, especially in TRECVid, we made a late fusion between our SIFT descriptor and ST-MP7EH. The idea was to show that whilst both descriptors provide good concept recognition separately, each one represents a different type of visual information: SIFT is a very accurate image (spatial) descriptor, while ST-MP7EH focuses more on the temporal. In our experiment we tried to prove that fusing two different descriptors in such a manner could significantly improve the MAP. Since we use the same learning technique for the 2 descriptors, we have one SVM "score" from SIFT and another for ST-MP7EH for each shot. These 2 scores can be fused using a linear combination, with the mix variable α as a free parameter. We computed for each shot $score_{fusion} = \alpha \cdot score_{ST-MP7EH} + (1-\alpha) \cdot score_{SIFT}$ for 10 values of α in the interval [0,1]. For each value of α we computed the ranked lists and calculated the MAP.

6 Results

6.1 Comparison with Spatio-Temporal Baseline Descriptors

Table 1 shows the APs and the MAP obtained using the ST-MP7EH descriptor compared to the 3 variants of our SIFT multi-keyframe baseline described in

5.2. Improvement is evident for concepts containing motion (either of the object, such as Boat or Bus, or the camera, as on Cityscape or Singing). Cityscape has an exceptionally high score because of the dominant vertical edges, which can be seen as a discriminative feature for the concept. We justify the low score of Airplane_flying by pointing out the spatial inconsistency (the object can be anywhere in the frame, at any scale, whereas our descriptor is not scale-invariant) and the lack of background information (the background is either an edge-less sky, or ground that is irrelevant to the concept).

Table 1. Comparison between ST-MP7EH and spatio-temporal baseline

Descriptor	ST-MP7EH	mkfSIFT1	mkfSIFT2	mkfSIFT3
Airplane_Flying	0.00021851	0.00589261	0.01793414	0.02514655
Boat_Ship	0.02262003	0.01880984	0.02367281	0.01797646
Bus	0.00187492	0.00348848	0.00514326	0.00629265
Cityscape	0.21769612	0.17755185	0.13681920	0.15426416
Classroom	0.00857810	0.00785321	0.00900273	0.00465120
Demonstration_Or_Protest	0.01465306	0.03806428	0.06042413	0.03266605
Hand	0.00342396	0.00335580	0.00572085	0.00759707
Nighttime	0.01671574	0.02773094	0.04030243	0.06348906
Singing	0.07864447	0.06039517	0.07210428	0.07564689
Telephones	0.00006563	0.00969003	0.00302432	0.00346670
MAP	0.03644900	0.03528322	0.03741482	0.03911968

6.2 Spatio-Temporal Performance Gain

Table 2. Comparison Between ST-MP7EH and MPEG-7 Edge Histogram

Descriptor	ST-MP7EH	MPEG-7 edge
Airplane_Flying	0.00041296	0.00210417
Boat_Ship	0.00437917	0.00527415
Bus	0.00039620	0.00006228
Cityscape	0.20201674	0.02011237
Classroom	0.02932826	0.00374221
Demonstration_Or_Protest	0.02609449	0.00629317
Hand	0.02561565	0.01700422
Nighttime	0.10152920	0.05869340
Singing	0.07730526	0.05053595
Telephones	0.00256774	0.00155078
MAP	0.04696450	0.01653720

This test has been performed on a subset of the training and test datasets, consisting on half (30307) the number of shots. The experiment compares MAP for ST-MP7EH and MPEG-7 Edge Histogram and shows how many concepts that are lacking in spatial recognition (i.e. Demonstration_Or_Protest) perform far better in spatio-temporal analysis. The results can be seen in table 2. Since ST-MP7EH actually uses Edge Histogram, the improvement is a measure of temporal relevance given by the concept. Note that the MAP for ST-MP7EH differs from the one in the previous experiment because of the different datasets used.

6.3 ST-MP7EH - SIFT Late Fusion

Figure 3 shows how different mixes between ST-MP7EH and SIFT perform. The 10 columns represent 10 values for the α parameter, from 0 (*pure* ST-MP7EH) to 1 (*pure* SIFT). The first observation is that individually ST-MP7EH has a higher MAP (0.046964567) than SIFT (0.029211185) for these datasets. The fusion shows that there is clearly an optimum for each concept where the MAP from the fusion exceeds both descriptors. This is the result of complementary information that ST-MP7EH and SIFT are able to describe. The average gain in precision attributable to latent fusion is of 18.86782%, corresponding to a MAP of 0.055825 for a common value of $\alpha = 0.43$. We can also pick an optimum α value for each concept, which gives an upper bound of improvement of 22.823%, or a MAP of 0.057683244.

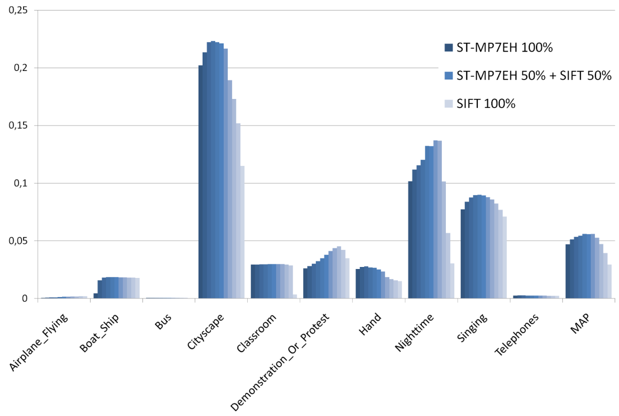


Fig. 3. Average Precision for late fusion between ST-MP7EH and SIFT

7 Conclusions

In this paper we presented a short overview of visual descriptors used in video retrieval concentrating on edge features, we proposed a novel spatio-temporal extension to the MPEG-7 Edge Histogram Descriptor, we described the computational method and provided experimental results of SVM-based retrieval in comparison to analogous spatio-temporal baseline descriptors, in comparison to its spatial predecessor and in fusion with SIFT.

Current large scale concept video retrieval systems show a slow adoption of dynamic features. In TRECVid, for example, only the Multimedia Event Detection task has provided significant research in spatio-temporal features. In the generalistic concept classifiers seen in the Semantic Indexing Task, the vast majority of systems still use only one keyframe to describe the entire shot. Only two different approaches have proven successful: local features (either in space or space-time) computed around STIP (spatio-temporal interest points) and the use of more than one keyframe in shot description. However these descriptors are part of complex systems where they participate in feature fusion. On all accounts, any method that analyzes the XYT volume is subject to a high processing and memory penalty. Our descriptor is invariant to changes in temporal scale, so that the frames of the shot can be subsampled up to a minimal rate. We can improve computation time by lowering the sampling rate with very little change in the feature vector.

Our experiment shows that MAP increases by almost 3 times (2.839) when we pass from the keyframe-based MPEG-7 Edge Histogram Descriptor to ST-MP7EH. This is remarkable since the two descriptors carry the same spatial information. The difference comes only from the temporal description via the usage of moments (mean and variance). We believe that these results should encourage the adoption of temporally-relevant description methods in such systems. The results of our late fusion experiment confirm that fusing a heterogeneous set of descriptors can yield a higher MAP thanks to the complementarity of the different representations. This is what we have been recently witnessing in TRECVid: large systems that collect information from color, texture, motion, audio, metadata features, etc. and perform a latent fusion similar to ours. To this end, EURECOM will use this descriptor in such a feature fusion scheme in the 2011 edition of TRECVid.

References

1. P. Over, G. Awad, J. Fiscus, B. Antonishek, and G. Qu, “TRECVID 2010 - an overview of the goals, tasks, data, evaluation mechanisms and metrics,” pp. 1–34, 2010.
2. C. G. M. Snoek and K. E. A. van de Sande, “The MediaMill TRECVID 2010 semantic video search engine,” in *Proceedings of the TRECVID Workshop*, 2010.
3. I. Laptev and T. Lindeberg, “Space-time interest points,” in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 432 –439 vol.1, oct. 2003.
4. J. Yang and A. G. Hauptmann, “(Un)Reliability of video concept detection,” in *Proceedings of the 2008 international conference on Content-based image and video retrieval*, CIVR ’08, (New York, NY, USA), pp. 85–94, ACM, 2008.

5. D. Weinland, R. Ronfard, and E. Boyer, “A survey of vision-based methods for action representation, segmentation and recognition,” *Comput. Vis. Image Underst.*, vol. 115, pp. 224–241, February 2011.
6. W. Ren, S. Singh, M. Singh, and Y. Zhu, “State-of-the-art on spatio-temporal information-based video retrieval,” *Pattern Recognition*, vol. 42, no. 2, pp. 267 – 282, 2009.
7. B. S. Manjunath, J. rainer Ohm, V. V. Vasudevan, and A. Yamada, “Color and texture descriptors,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, pp. 703–715, 1998.
8. Y. Shimoda, A. Noguchi, and K. Yanai, “UEC at TRECVID 2010 semantic indexing task.”
9. M. Naito, K. Hoashi, K. Matsumoto, M. Shishibori, K. Kita, A. Kutics, A. Nakagawa, F. Sugaya, and Y. Nakajima, “High-level feature extraction experiments for TRECVID 2007,” in *TRECVID'07*, 2007.
10. S. Tang, Y. dong Zhang, J. tao Li, X. feng Pan, T. Xia, M. Li, A. Liu, L. Bao, S. chang Liu, Q. feng Yan, and L. Tan, “Rushes Exploitation 2006 By CAS MCG.”
11. A. Moumtzidou, A. Dimou, N. Gkalelis, S. Vrochidis, V. Mezaris, and I. Kompatsiaris, “ITI-CERTH participation to TRECVID 2010,” 2010.
12. M. Rautiainen, M. Varanka, I. Hanski, M. Hosio, A. Pramila, J. Liu, and T. Ojala, “TRECVID 2005 Experiments at MediaTeam Oulu,” 2005.
13. F. Mahmoudi, J. Shanbehzadeh, A.-M. Eftekhari-Moghadam, and H. Soltanian-Zadeh, “Image retrieval based on shape similarity by edge orientation autocorrelogram,” *Pattern Recognition*, vol. 36, no. 8, pp. 1725 – 1736, 2003.
14. M. Yang, S. Ji, W. Xu, J. Wang, F. Lv, K. Yu, Y. Gong, M. Dikmen, D. J. Lin, and T. S. Huang, “Detecting Human Actions in Surveillance Videos.”
15. D. K. Park, Y. S. Jeon, and C. S. Won, “Efficient use of local edge histogram descriptor,” in *Proceedings of the 2000 ACM workshops on Multimedia*, MULTIMEDIA '00, (New York, NY, USA), pp. 51–54, ACM, 2000.
16. P. Scovanner, S. Ali, and M. Shah, “A 3-dimensional sift descriptor and its application to action recognition,” in *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, (New York, NY, USA), pp. 357–360, ACM, 2007.
17. A. Kläser, M. Marszalek, and C. Schmid, “A spatio-temporal descriptor based on 3d-gradients,” in *In BMVC 08*.
18. A. Efros, A. Berg, G. Mori, and J. Malik, “Recognizing action at a distance,” in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 726 –733 vol.2, oct. 2003.
19. R. Polana and R. Nelson, “Recognition of motion from temporal texture,” in *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, pp. 129 –134, jun 1992.
20. T. Darrell and A. Pentland, “Space-time gestures,” in *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR '93., 1993 IEEE Computer Society Conference on*, pp. 335 –340, jun 1993.
21. F. Liu and R. Picard, “Finding periodicity in space and time,” in *Computer Vision, 1998. Sixth International Conference on*, pp. 376 –383, jan 1998.
22. C.-C. Chang and C.-J. Lin, “LIBSVM: a Library for Support Vector Machines,” 2001.