Proc. IEEE International Conference on Image Processing (ICIP 2010), Hong Kong, China, September 2010, pp. 4697-4700.

ON THE USE OF FEATURE TRACKS FOR DYNAMIC CONCEPT DETECTION IN VIDEO

Vasileios Mezaris, Anastasios Dimou, Ioannis Kompatsiaris

Informatics and Telematics Institute / Centre for Research and Technology Hellas
6th Km Charilaou-Thermi Road, Thermi 57001, Greece
{bmezaris, dimou, ikom}@iti.gr

ABSTRACT

This paper proposes the use of feature tracks for the detection of concepts in video, particularly dynamic concepts. Feature tracks are defined as sets of local interest points found in different frames of a video shot that exhibit spatio-temporal and visual continuity, defining a trajectory in the 2D+Time space. The extraction of feature tracks and the selection and representation of an appropriate subset of them allow the generation of a Bag-of-Spatiotemporal-Words model for the shot, which facilitates capturing the dynamics of video content. The experimental evaluation of the proposed approach highlights how the selection of such feature tracks for the definition of the Bag-of-Spatiotemporal-Words model enhances the results of traditional keyframe-based concept detection techniques.

Index Terms— Video signal processing, image sequence analysis, motion analysis, concept detection

1. INTRODUCTION

The development of algorithms for the automatic understanding of the semantics of multimedia and in particular of video content is currently one of the major challenges in multimedia research. This is motivated by the ever-increasing pace at which video content is generated, rendering any annotation scheme that requires human labor unrealistically expensive and unpractical for use in the majority of potential applications.

This work focuses on the detection of high-level concepts in video, particularly dynamic ones (e.g. action- or motion-related concepts, as opposed to static ones). It builds upon previous work on local interest point detection and description to propose the extraction, selection and representation of Feature Tracks. These features compactly describe the appearance and the long-term motion of local regions and are invariant, among others, to camera motion, in contrast to both 2D interest point descriptors and their known extensions to spatio-temporal interest points. The proposed feature tracks are shown to be suitable for the generation of a Bag-of-Spatiotemporal-Words (BoSW) model that facilitates concept detection in video.

The rest of the paper is organized as follows: in section 2, previous work on local interest point detection and description is discussed. In section 3 feature track extraction and selection are presented, while the representation of feature tracks using the LIFT descriptor and their use in a Bag-of-Spatiotemporal-Words model are discussed in section 4. Experimental results are reported in section 5 and finally conclusions are drawn in section 6.

This work was supported by the European Commission under contracts FP6-045547 VIDI-Video and FP7-248984 GLOCAL.

2. RELATED WORK

Several approaches to scale-invariant interest point detection and description in still images have been proposed and are widely used in still image understanding tasks (image classification, object detection, etc.) as well as other applications. These include SIFT [1], SURF [2], and techniques introducing color information to the original grey-value SIFT [3]. For the application of high-level feature extraction in generic image collections, the above descriptors are typically used to build a Bag-of-Words (BoW) model [4], which involves the definition of a "vocabulary" of visual words and the subsequent representation of each image as the histogram of the visual words (i.e. corresponding interest points) found in it.

Large scale video analysis for the purpose of high-level feature extraction, using local invariant features, is in most cases performed at the key-frame level [5]. Thus, the video analysis task reduces to still image analysis. This has obvious advantages in terms of computational complexity, but on the other hand completely disregards the temporal dimension of video and the wealth of information that is embodied in the evolution of the visual signal along time. The latter, particularly long-term trajectories (e.g. [6]), is generally considered to be very important for video analysis. For concept detection, similarly to other analysis tasks, the use of video data in excess of one single key-frame (e.g. multiple key-frames per shot [7]) has been shown to lead to improved results.

In order to introduce temporal information in the interest-point-based representation of video shots, the use of spatio-temporal (as opposed to spatial-only) interest point detectors has been proposed [8]. Spatio-temporal interest points are defined as locations in the video where intensity values present significant variations both in space and in time. In [9] and other works, such points are used for human action categorization, since the abrupt changes in motion that trigger the detection of spatio-temporal interest points can be useful in discriminating between different classes of human activity (walking, jumping, etc.). However, spatio-temporal interest points define 3D volumes in the video data that typically neither account for possible camera motion nor capture long-term local region trajectories. To alleviate these drawbacks, the tracking of spatial interest points across successive frames has been proposed for applications such as object tracking [10] and visualization of pedestrian traffic flow in surveillance video [11]. In [12], the problem of object mining in video is addressed by tracking SIFT features and subsequently clustering them, to identify differently moving objects within a shot. In [13, 14] interest points are tracked and either the motion information alone [13] or appearance and motion information in separate BoW models [14] are used for action recognition in video. However, neither one of the previous works on tracking spatial interest points [10]-[14] uses the outcome of tracking for defining a BoSW model of the shot, as in the present work.

3. FEATURE TRACKS

3.1. Feature Track extraction

Let S be a video shot comprising T frames, $S = \{I_t\}_{t=0}^{T-1}$. Application of interest point detection and description techniques (e.g. [1, 2, 3]) on any frame I_t of S results in the extraction of a set of interest point descriptions $\Phi_t = \{\phi_m\}_{m=1}^{M_t}$, where interest point ϕ_m is defined as $\phi_m = [\phi_m^x, \phi_m^y, \phi_m^d]$. ϕ_m^x, ϕ_m^y denote the coordinates of the corresponding local region's centroid on the image grid and ϕ_m^d is the local descriptor vector, e.g. an 128-element SIFT vector. In this work, the SIFT method was used for interest point detection and description, due to its well-documented [1, 15] invariance properties.

Having detected and described interest points in all frames of S, a temporal correspondence between an interest point $\phi_m \in \Phi_t$ and one interest point of the previous frame can be established by local search in a square spatial window of dimension $2 \cdot \sigma + 1$ of frame I_{t-1}, i.e. by examining if one or more $\phi_n \in \Phi_{t-1}$ exist that satisfy:

$$|\phi_m^x - \phi_n^x| \leq \sigma, \;\; |\phi_m^y - \phi_n^y| \leq \sigma, \;\; d(\phi_m^d, \phi_n^d) \leq d_{sim} \quad (1)$$

where d(.,.) is the Euclidean distance. The latter is chosen for consistency with the K-Means clustering used in section 4.2 for assigning the extracted tracks to Words of the BoSW model. If multiple interest points satisfying Eq. (1) exist, the one for which quantity $d(\phi_m^d, \phi_n^d)$ is minimized is retained. When such an interest point ϕ_n exists, the interest point $\phi_m \in \Phi_t$ is appended to the feature track where the former belongs, while otherwise (as well as when processing the first frame of the shot) the interest point ϕ_m is considered to be the first element of a new feature track.

Repeating the temporal correspondence evaluation for all interest points and all pairs of consecutive frames in S results in the extraction of a set Ψ of feature tracks, $\Psi = \{\psi_k\}_{k=1}^{K}$, where $\psi_k = [\psi_k^x, \psi_k^y, \psi_k^d]$. ψ_k^d is the average descriptor vector of a feature track, estimated by element-wise averaging of all interest point descriptor vectors ϕ_m^d of the feature track as in [12], while ψ_k^x is the corresponding time series of camera-motion-compensated interest point displacement in the x-axis between successive frames of S where the feature track is present, and similarly ψ_k^y for the y-axis. Thus, $\xi_k = [\psi_k^x, \psi_k^y]$ is the long-term trajectory of the interest point that generates the feature track: $\psi_k^x = [\psi_k^{x,t_{k1}}, \; \psi_k^{x,t_{k1}+1}, \; ... \; \psi_k^{x,t_{k2}}]$ where $t_{k2} > t_{k1}$ (and similarly for ψ_k^y). The values $\psi_k^{x,t}$ are estimated for any given t by initially using the differences $\phi_m^x - \phi_n^x$, $\phi_m^y - \phi_n^y$ for all identified valid pairs of interest points between frames I_t, I_{t-1} to form a sparse, non-regular motion field for the corresponding pair of frames; subsequently, the 8 parameters of the bilinear motion model, representing the camera motion, are estimated from this field using least-squares estimation and an iterative rejection scheme, as in [6]. Then $\psi_k^{x,t}$ and $\psi_k^{y,t}$ are eventually calculated as the differences between the initial displacement of the corresponding interest point's centroid between times $t-1$ and t, and the estimated camera motion at the location of the centroid.

The objective of estimating the camera motion in the above process is to ensure that the extracted feature tracks are invariant to camera motion. The latter (i.e. camera motion) may also be an important cue, but should probably be described separately for the entire shot; if it was encoded in every feature track, not only would there be duplication of information, we would also be unable to differentiate between object and camera motion in the feature tracks' description.

3.2. Feature Track selection

The described feature track extraction process typically results in the extraction of a large number of feature tracks (e.g. in the order of tens of thousands) for every shot. These exhibit significant differences in their temporal duration, with the track length $t_{k2} - t_{k1}$ ranging from 0 to $T-1$, T being the number of frames in the shot (Fig. 1). Besides the practical problems associated with storing and using such a large number of descriptors for every shot, the possible presence of noisy or otherwise erroneous tracks among those originally extracted may adversely affect concept detection. Therefore, selecting a suitable subset of these feature tracks is proposed.

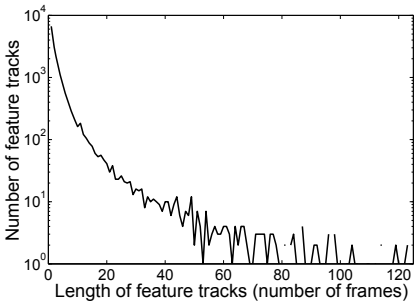

Fig. 1. Example of the distribution of feature tracks extracted for a shot according to their temporal duration.

One possible criterion for selecting a subset of feature tracks is their repeatability under variations (e.g. perspective, scale, and illumination variations). Repeatability is among the main requirements for any descriptor. In this work, it is proposed that the repeatability of a track is approximated by examining the temporal duration of it. More specifically, let us assume that R denotes the real-world scene that is depicted in shot S. Under constant illumination conditions and assuming no local (object) motion, the result of capturing scene R with an ideal static camera would be an ideal image I_r. Then, every image $I_t \in S$ can be seen as a different noisy observation of I_r, affected by image acquisition noise and possible global and local motion as well as perspective, scale, and illumination variations. Similarly, every interest point in image I_t that is part of an extracted feature track ψ_k can be perceived as the (successful) result of detecting the corresponding ideal interest point of I_r under the specific variations affecting image I_t. Consequently, the probability of a specific feature track being present in one frame of S can be used as a measure of the repeatability of the interest point that defines this feature track, thus also as a measure of the relevant repeatability of the feature track itself in comparison to other feature tracks of the shot.

Following this discussion, in this work the probability of a specific feature track being present in one frame of S is calculated as the number of frames in which the track extends, divided by the total number of frames of the shot,

$$p(\psi_k) = \frac{t_{k2} - t_{k1}}{T - 1}, \quad (2)$$

and is used as a measure of the feature track's repeatability. Consequently, the feature tracks of set Ψ generated for shot S are ordered according to $p(\psi_k)$ (equivalently, in practice, according to $t_{k2} - t_{k1}$) in descending order and the N first tracks are selected for generating the BoSW model of the shot. This track selection strategy is evaluated against two others in the experimental results section.

It should be emphasized that repeatability is just one possible criterion for selecting feature tracks, and the most repeatable features are not necessarily the most informative ones as well; thus, jointly considering repeatability and additional criteria may be beneficial.

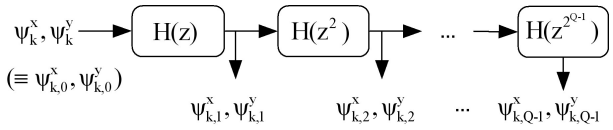

Fig. 2. Filter bank used for capturing motion at different time-scales.

4. BAG-OF-SPATIOTEMPORAL-WORDS

4.1. Feature Track representation

The selected feature tracks are variable-length feature vectors, since the number of elements comprising ψ_k^x and ψ_k^y is proportional to the number of frames that the feature was successfully tracked in. This fact, together with other possible track artefacts (e.g. the extraction of partial tracks, due to failure in interest point matching between consecutive frames, occlusions etc.) make the matching of feature tracks non-trivial and render their current representation unsuitable for direct use in a BoW-type approach. To this end, each motion trajectory is transformed to a fixed-length descriptor vector that attempts to capture the most important characteristics of the motion.

To capture motion at different time-scales, ψ_k^x and ψ_k^y are initially subject to low-pass filtering using a filter bank shown in Fig. 2, based on the lowpass Haar filter $H(z) = \frac{1}{2}(1 + z^{-1})$. This results in the generation of a family of trajectories, $\xi_{k,q} = [\psi_{k,q}^x, \psi_{k,q}^y]$, $q = 0, ..., Q-1$, as shown in Fig. 2, which due to the simplicity of the Haar filter are conveniently calculated as follows:

$$\psi_{k,q}^x = [\psi_{k,q}^{x,t_{k1}+2^q-1},\ \psi_{k,q}^{x,t_{k1}+2^q},\ ...\ \psi_{k,q}^{x,t_{k2}}] \tag{3}$$

$$\psi_{k,q}^{x,t} = \frac{1}{2^q}\sum_{i=0}^{2^q-1}\psi_k^{x,t-i} \tag{4}$$

The y-axis elements of the trajectory are calculated similarly.

For any trajectory $\xi_{k,q}$, the histogram of motion directions at granularity level θ is defined as a histogram of $\frac{\pi}{\theta}$ bins: $[0, \theta)$, $[\theta, 2 \cdot \theta)$,..., $[\pi - \theta, \pi)$. When $\pi \leq \theta < 2 \cdot \pi$, $\theta' = \theta - \pi$ is used instead of θ for assigning the corresponding elementary motion to the appropriate bin of the histogram. The value of each bin is defined as the number of elementary motions $[\psi_{k,q}^{x,t},\ \psi_{k,q}^{y,t}]$ of the trajectory that fall into it, normalized by division with the overall number of such elementary motions that belong to the examined trajectory. $\lambda(\xi_{k,q}, \theta)$ is defined as the vector of all bin values for a given $\xi_{k,q}$ and a constant θ.

Then, the initial trajectory ξ_k can be represented across different time-scales as a fixed length vector μ_k,

$$\begin{aligned}\mu_k = \Big[&\lambda(\xi_{k,0}, \frac{\pi}{2}),\ \ \lambda(\xi_{k,1}, \frac{\pi}{2}),\ ...\ \lambda(\xi_{k,Q-1}, \frac{\pi}{2})\\ &\lambda(\xi_{k,0}, \frac{\pi}{4}),\ \ \lambda(\xi_{k,1}, \frac{\pi}{4}),\ ...\ \lambda(\xi_{k,Q-1}, \frac{\pi}{4}),\ ...\\ &\lambda(\xi_{k,0}, \frac{\pi}{2^J}),\ \ \lambda(\xi_{k,1}, \frac{\pi}{2^J}),\ ...\ \lambda(\xi_{k,Q-1}, \frac{\pi}{2^J})\Big]\end{aligned} \tag{5}$$

and the corresponding Local Invariant Feature Track (LIFT) descriptor is defined as

$$LIFT(\psi_k) = [\psi_k^d,\ \ \mu_k] \tag{6}$$

This descriptor is invariant to scale and camera motion, but not to the orientation of local motion, since the latter is considered to be an important cue for dynamic concept detection.

4.2. Shot representation

The LIFT descriptors of the feature tracks extracted and selected according to the processes of section 3 for a video shot can be used for generating a Bag-of-Spatiotemporal-Words (BoSW) model. This will essentially describe the shot in terms of classes of "similarly-moving, visually-similar local regions", rather that simply "visually-similar local regions" (detected by either spatial or spatio-temporal interest point detectors), as in the current state-of-the-art, e.g. [7, 9]. The BoSW model is expected to allow for the improved detection of dynamic concepts in video, in contrast to the traditional keyframe-based BoW that by definition targets the detection of static concepts. Furthermore, since the shot features used in the BoW and BoSW models are different and, to some degree, complementary, it is expected that combining the two models can result in further improvement of the detection rates for both dynamic and static concepts.

For the generation of the BoSW model, the typical process of generating BoW descriptions from any set of local descriptors is followed. Thus, K-Means clustering using a fixed number of clusters is performed on a large collection of LIFT descriptors for initially identifying a set of Words (i.e. the centroids of the clusters). Hard- or soft-assignment of each one of the LIFTs of a given shot to these words can then be performed for estimating the histogram that represents a given shot on the basis of the defined spatio-temporal words.

5. EXPERIMENTAL RESULTS

In the experimental evaluation of the proposed techniques, the TRECVID 2007 training and test datasets were employed (comprising 50 hours of video each, and 18120 and 18142 shots respectively) together with the 20 concepts that were defined on this dataset for the TRECVID 2009 contest[1]. In extracting the feature tracks, parameter σ defining the local window where correspondences between SIFT descriptors are evaluated was set to 20, and parameter d_{sim} used for evaluating the similarity of SIFT descriptors in different frames was set to 40000. Using four different timescales ($Q = 4$) and three granularity levels θ (i.e. $J = 3$ in Eq. (5)) for representing the trajectory information of the extracted feature tracks resulted in the LIFT descriptor of each feature track being a 184-element vector.

A first series of experiments was carried out to evaluate the appropriate number of feature tracks that should be used for representing each shot, given the above feature track extraction and representation parameter choices. A BoSW model using hard assignment and 500 words was used to this end, together with standard Support Vector Machine classifiers. It should be noted that this is only a baseline configuration; it is used for efficiently evaluating certain characteristics of the proposed BoSW, and is neither optimal not in par with SoA works such as [7], where 4000 words, soft assignment, multiple color SIFT variants, and additional techniques such as pyramidal decomposition are combined, increasing the dimension of the vector representing each shot from 500 (as in our baseline configuration) to about 100000. The results (average precision@2000 [5]) are shown in Fig. 3(a), where it can be seen that using 2500 feature tracks per shot leads to the best results overall.

A second series of experiments was carried out to evaluate the soundness of the feature track selection process of section 3.2. Specifically, the selection of the 2500 tracks with the highest probability $p(\psi_k)$, as proposed in section 3.2 (denoted as selection criterion "BB" in the sequel) was compared with a) the selection of the 2500 tracks with the highest probability $p(\psi_k)$ after removing

[1] http://www-nlpir.nist.gov/projects/trecvid/

from set Ψ those feature tracks used by selection criterion "BB" (denoted as "SB" in the sequel), and b) the random selection of 2500 feature tracks from set Ψ (selection criterion "RR"). The LIFT descriptor was used in all the above cases for representing the selected tracks and for forming a 500-word BoSW model. Experimentation with the 500-word keyframe-based BoW model that uses SIFT descriptors was also carried out, for comparing BoSW and BoW when used in isolation. The results (Fig. 3(b)) show that selection criterion "BB" significantly outperforms criteria "SB" and "RR". The BoSW model using selection criterion "BB" by itself performs comparably to the keyframe-based BoW model overall, but considerably better when considering only dynamic concepts.

In a third series of experiments, the merit of combining the BoSW and BoW models was evaluated. The combination of the two was performed by concatenating the shot descriptions produced by each of them, similarly to how different BoW models based on different color SIFT variants are combined in [7]. In Table 1, BoW and the combination of BoW and BoSW (using selection criterion "BB") are compared using a) the baseline configuration used in the previous experiments: 500 words and hard assignment, and b) 500 words, soft assignment, a spatial pyramid of 2 levels for BoW and, in a similar fashion, a temporal pyramid for BoSW. Additionally, in the latter case 5 granularity levels θ (i.e. $J = 5$ in Eq. (5)), instead of 3, are used. The results of Table 1 document the contribution of the proposed BoSW model to improved performance when combined with the BoW model, compared to the latter alone, as well as the applicability of techniques such as soft assignment and pyramidal decomposition (particularly temporal pyramids) to BoSW.

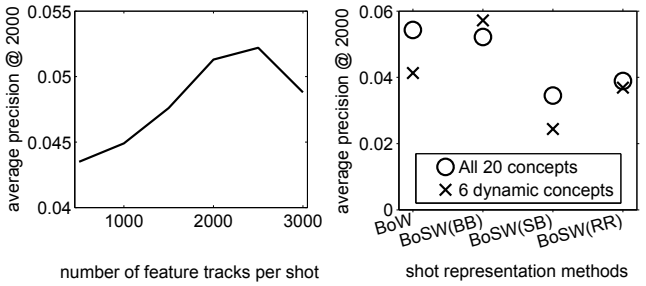

Fig. 3. Evaluation of a) the impact of the number of feature tracks used for representing each shot, and b) the impact of different shot representation techniques, on concept detection performance.

Table 1. Comparison between BoW, combination of BoW and BoSW (average precision@2000 for all 20 / 6 dynamic concepts).

	BoW		BoW+BoSW(BB)	
considered concepts:	20	6	20	6
500 words, hard assign.	0.054	0.041	0.068	0.056
500 words, soft assign., pyramidal decomp.	0.084	0.088	0.102	0.113

6. CONCLUSIONS

In this work the use of feature tracks was proposed for jointly capturing the spatial attributes and the long-term motion of local regions in video, and in particular techniques for the extraction, selection, representation and use of feature tracks for constructing a Bag-of-Spatiotemporal-Words model for the video shots were presented. Experimental evaluation of the proposed approach on the corpus of TRECVID 2007 revealed its potential for concept detection in video, particularly when considering dynamic rather than static concepts.

7. REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. of Computer Vision*, vol. 60, pp. 91–110, 2004.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[3] G. J. Burghouts and J.-M. Geusebroek, "Performance Evaluation of Local Colour Invariants," *Computer Vision and Image Understanding*, vol. 113, pp. 48–62, 2009.

[4] C. Dance, J. Willamowski, L.X. Fan et.al., "Visual categorization with bags of keypoints," in *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, Prague, CZ, May 2004.

[5] A. F. Smeaton, P. Over, and W. Kraaij, "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements," in *Multimedia Content Analysis, Theory and Applications*, A. Divakaran, Ed., pp. 151–174. Springer, 2009.

[6] V. Mezaris, I. Kompatsiaris, N.V. Boulgouris, and M.G. Strintzis, "Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval," *IEEE Trans. on CSVT*, vol. 14, no. 5, pp. 606–621, May 2004.

[7] C.G.M. Snoek, K.E.A. van de Sande, O. de Rooij et.al., "The MediaMill TRECVID 2008 Semantic Video Search Engine," in *Proc. TRECVID 2008 Workshop*, USA, Nov. 2008.

[8] I. Laptev, "On Space-Time Interest Points," *Int. J. of Computer Vision*, vol. 64, no. 2/3, pp. 107–123, 2005.

[9] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *Int. J. of Computer Vision*, vol. 79, no. 3, pp. 299–318, Sept. 2008.

[10] H. Zhou, Y. Yuan, and C. Shi, "Object tracking using SIFT features and mean shift," *Computer Vision and Image Understanding*, vol. 113, pp. 345–352, 2009.

[11] Y. Tsuduki and H. Fujiyoshi, "A Method for Visualizing Pedestrian Traffic Flow Using SIFT Feature Point Tracking," in *Proc. 3rd Pacific-Rim Symposium on Image and Video Technology*, Tokyo, Japan, Jan. 2009.

[12] A. Anjulan and N. Canagarajah, "A Unified Framework for Object Retrieval and Mining," *IEEE Trans. on CSVT*, vol. 19, no. 1, pp. 63–76, Jan. 2009.

[13] N. Moenne-Loccoz, E. Bruno, and S. Marchand-Maillet, "Local Feature Trajectories for Efficient Event-Based Indexing of Video Sequences," in *Proc. CIVR*, Tempe, USA, July 2006.

[14] J. Sun, X. Wu, S. Yan et.al., "Hierarchical Spatio-Temporal Context Modeling for Action Recognition," in *Proc. CVPR*, Miami, USA, June 2009.

[15] K. Mikolajczyk and C. Schmid, "Performance Evaluation of Local Descriptors," *IEEE Trans. on PAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.