V. Mezaris, A. Dimou, I. Kompatsiaris, "Local Invariant Feature Tracks for High-Level Video Feature Extraction", Analysis, Retrieval and Delivery of Multimedia Content, N. Adami, A. Cavallaro, R. Leonardi, P. Migliorati (Eds.), Springer Lecture Notes in Electrical Engineering vol. 158, ISBN 978-1-4614-3830-4, 2012. The original publication is available at www.springerlink.com.

Local Invariant Feature Tracks for High-Level Video Feature Extraction

Vasileios Mezaris, Anastasios Dimou, and Ioannis Kompatsiaris

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

Abstract. In this work the use of feature tracks for the detection of high-level features (concepts) in video is proposed. Extending previous work on local interest point detection and description in images, feature tracks are defined as sets of local interest points that are found in different frames of a video shot and exhibit spatio-temporal and visual continuity, thus defining a trajectory in the 2D+Time space. These tracks jointly capture the spatial attributes of 2D local regions and their corresponding long-term motion. 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. Experimental evaluation of the proposed approach on two challenging datasets (TRECVID 2007, TRECVID 2010) highlights how the selection, representation and use of such feature tracks enhances the results of traditional keyframe-based concept detection techniques.

Keywords: Feature tracks, video concept detection, trajectory, LIFT descriptor, Bag-of-Spatiotemporal-Words

1 Introduction

The development of algorithms for the automatic understanding of the semantics of multimedia and in particular of video content, and the semantic indexing by means of high-level features (concepts) corresponding to semantic classes (objects, events) 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 on anything but a very restricted subset of the generated content, which may be of unusually high value or importance (e.g. cinema productions, medical content).

Research efforts towards the goal of high-level video feature extraction have followed in the last decade or so several different directions that have the potential to contribute to this goal, ranging from temporal or spatio-temporal segmentation [1, 2] to key-frame extraction, video content representation using global shot or image features, local interest point detection and description [3], creation of visual lexicons for video representation (Bag-of-Words [4]), machine learning for associating low-level and high-level features, etc. Typically, techniques belonging to several of the aforementioned categories need to be carefully combined for extracting high-level video features. The latter are useful in a wide variety of media organization and analysis tasks, including interactive retrieval and the detection of scenes and high-level events in video [5, 6].

This work focuses on video content representation, and in particular 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 capturing the dynamics of video content, allowing the more reliable detection of high-level features that have a strong temporal dimension (e.g. "people-dancing").

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

V. Mezaris, A. Dimou, I. Kompatsiaris, "Local Invariant Feature Tracks for High-Level Video Feature Extraction", Analysis, Retrieval and Delivery of Multimedia Content, N. Adami, A. Cavallaro, R. Leonardi, P. Migliorati (Eds.), Springer Lecture Notes in Electrical Engineering vol. 158, ISBN 978-1-4614-3830-4, 2012. The original publication is available at www.springerlink.com.

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 in other applications. SIFT [3] is probably the most widely adopted method; SIFT-based descriptors are shown in [7] to outperform several previously proposed techniques for local region description. More recent work on this topic includes SURF [8], which focuses mostly on speeding-up the interest point detection and description process, and [9], which examines the introduction of color information to the original grey-value SIFT. 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 (typically, created by clustering the interest point descriptors coming from a large number of images and then selecting the resulting centroids as 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 features, is in most cases performed at the key-frame level [10]. 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 video frames along time. The temporal evolution of the video signal, i.e. motion, is generally considered to convey very important information in video, being a key element of several video understanding and manipulation tasks, e.g. retrieval [11]. Long-term region trajectories in particular, rather than the motion at the frame level, have been shown to be very useful for video segmentation, indexing and retrieval in several works (e.g. [1]). Similarly to other analysis tasks, the use of video data in excess of one single key-frame (e.g. using multiple key-frames per shot [12], or treating all frames as key-frames and also considering their temporal succession [13]) for high-level feature extraction has been shown to lead to improved results.

In order to introduce temporal information in the interest-point-based representation of video shots, in [14] spatial interest points are detected using the SIFT methodology and additional motion constraints; the detected points are described using both visual and motion information. In [15], the use of spatio-temporal (as opposed to spatial-only) interest point detectors is proposed. Spatio-temporal interest points are defined as locations in the video where intensity values present significant variations both in space and in time. In [16] 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 [17] and the visualization of pedestrian traffic flow in surveillance video [18]. In [19], 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 [20, 21], interest points are tracked and either the motion information alone [20] or appearance and motion information in separate BoW models [21] are used for action recognition in video. However, neither one of the previous works on tracking spatial interest points 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 shot comprising T frames, $S = \{I_t\}_{t=0}^{T-1}$, coming from the temporal sub-sampling of the original video shot $S^0 = \{I_\tau\}_{\tau=0}^{T^0-1}$ by a factor of a; $T = \lceil T^0/a \rceil$.

Application of one of the available combinations of interest point detection and description techniques (e.g. [3, 8, 9]) 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 M_t is the total number of interest points detected in the frame, and 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 [3, 7] 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

V. Mezaris, A. Dimou, I. Kompatsiaris, "Local Invariant Feature Tracks for High-Level Video Feature Extraction", Analysis, Retrieval and Delivery of Multimedia Content, N. Adami, A. Cavallaro, R. Leonardi, P. Migliorati (Eds.), Springer Lecture Notes in Electrical Engineering vol. 158, ISBN 978-1-4614-3830-4, 2012. The original publication is available at www.springerlink.com.

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 the following conditions:

$$|\phi_m^x - \phi_n^x| \leq \sigma \ , \tag{1}$$

$$|\phi_m^y - \phi_n^y| \leq \sigma \ , \tag{2}$$

$$d(\phi_m^d, \phi_n^d) \leq d_{sim} \ , \tag{3}$$

where σ is a constant whose value is chosen such that a reasonably-sized square spatial window is considered during local search, and d(.,.) is the Euclidean distance. The latter was also used in [3] for keypoint matching across different images, and is chosen in this work for consistency with the K-Means clustering that is used at a later stage for assigning the extracted tracks to words of the BoSW model (Sect. 4.3). If multiple interest points satisfying (1)-(3) 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 [19], 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 in which the feature track is present. ψ_k^y is defined similarly 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 [1]. 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 simple interest point matching between successive frames of S, which is used as part of the proposed feature track extraction process, was chosen primarily for its simplicity; more elaborate techniques for tracking across frames have been proposed (e.g. [18]) and can be used instead, for producing more accurate feature tracks, if the added computational complexity is not a limiting factor. An example of the feature tracks that are extracted by the proposed procedure is shown in Fig. 1.

3.2 Feature Track Selection

The feature track extraction process, described in the previous section, 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. 2). 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.

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 hypothesized that the repeatability of a track can be 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 result of detecting the corresponding ideal interest point of I_r under the specific variations affecting image I_t. Of course, the assumption made here is that the correspondences established with the use of (1)-(3) are not erroneous. Consequently, the probability of a specific feature track being present in

V. Mezaris, A. Dimou, I. Kompatsiaris, "Local Invariant Feature Tracks for High-Level Video Feature Extraction", Analysis, Retrieval and Delivery of Multimedia Content, N. Adami, A. Cavallaro, R. Leonardi, P. Migliorati (Eds.), Springer Lecture Notes in Electrical Engineering vol. 158, ISBN 978-1-4614-3830-4, 2012. The original publication is available at www.springerlink.com.

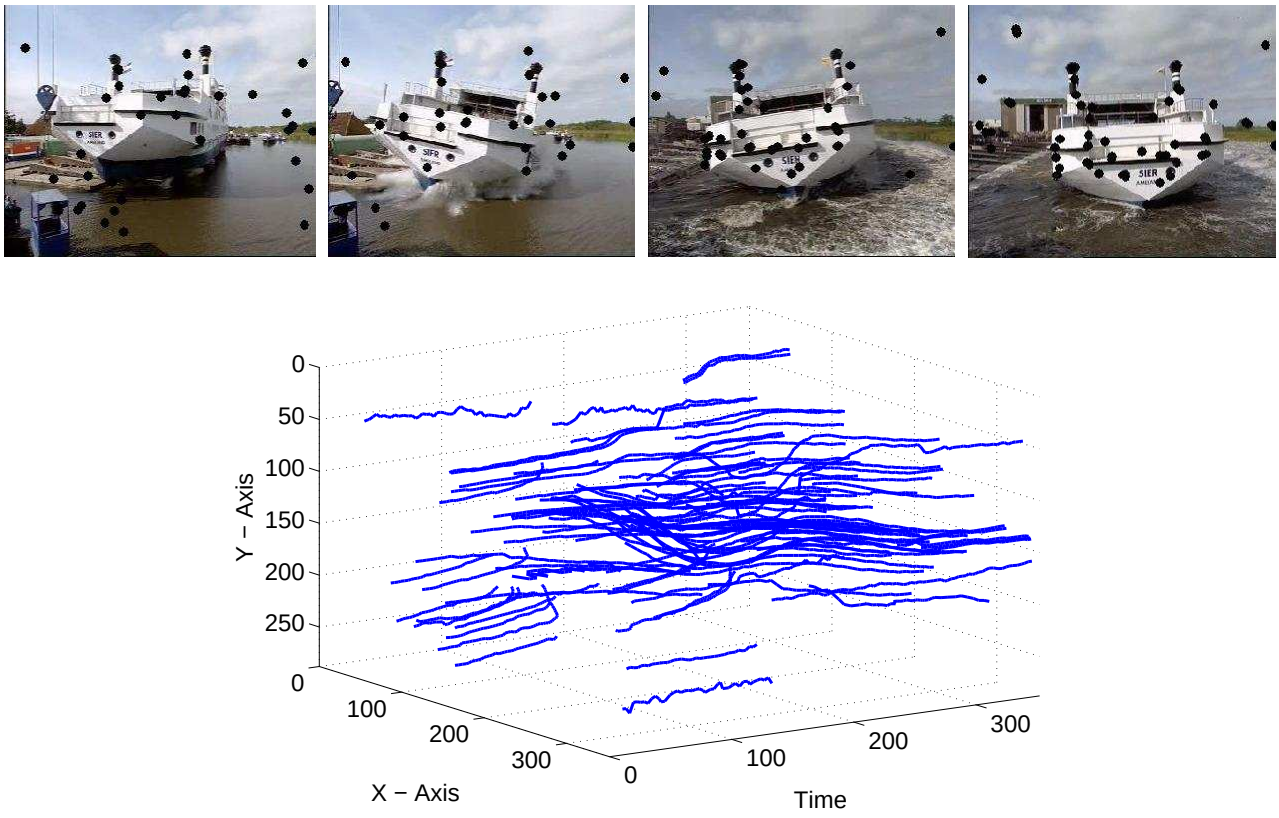

Fig. 1. Example of a few interest points that belong to extracted feature tracks (marked on four indicative frames of a shot), and an overview of the corresponding feature tracks in the 2D+Time space for the whole shot.

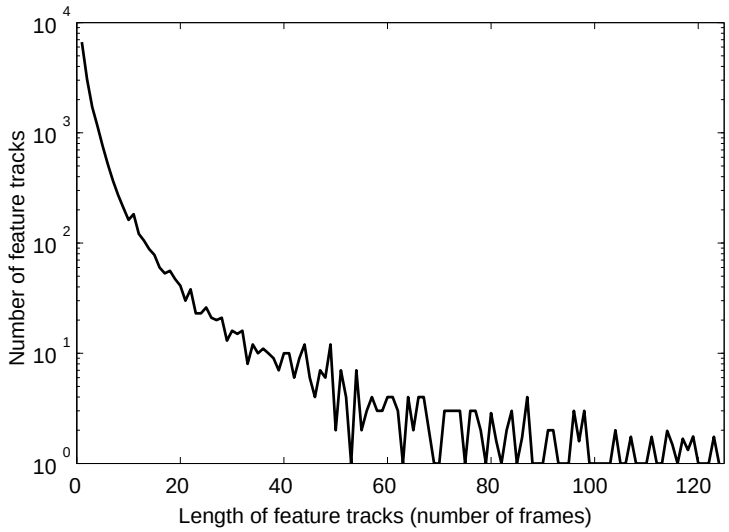

Fig. 2. Example of the distribution of feature tracks extracted for a shot, according to their temporal duration.

one frame of S can be used as a measure of the repeatability of the interest point that defines this feature track,

V. Mezaris, A. Dimou, I. Kompatsiaris, "Local Invariant Feature Tracks for High-Level Video Feature Extraction", Analysis, Retrieval and Delivery of Multimedia Content, N. Adami, A. Cavallaro, R. Leonardi, P. Migliorati (Eds.), Springer Lecture Notes in Electrical Engineering vol. 158, ISBN 978-1-4614-3830-4, 2012. The original publication is available at www.springerlink.com.

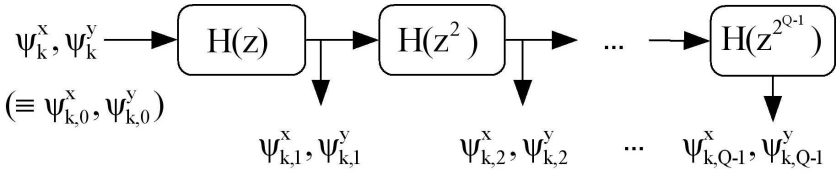

Fig. 3. Filter bank used for capturing motion at different time-scales.

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,

$$p(\psi_k) = \frac{t_{k2} - t_{k1}}{T - 1} \ , \tag{4}$$

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.

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. Furthermore, note that the temporal duration of a track being a good approximation of its repeatability is only a hypothesis that we make; this needs to be experimentally verified. To this end, the track selection strategy described above, which is based on this hypothesis, is evaluated against two other possible such strategies in the experimental results section.

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^x_k and ψ^y_k is proportional to the number of frames that the feature was 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. For this reason, 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^x_k and ψ^y_k are initially subject to low-pass filtering using a filter bank shown in Fig. 3, based on the lowpass Haar filter $H(z) = \frac{1}{2}(1 + z^{-1})$. This results in the generation of a family of trajectories, $\xi_{k,q} = [\psi^x_{k,q}, \psi^y_{k,q}]$, $q = 0, ..., Q-1$, as shown in Fig. 3, which due to the simplicity of the Haar filter are conveniently calculated as follows:

$$\psi^x_{k,q} = [\psi^{x,t_{k1}+2^q-1}_{k,q}, \ \psi^{x,t_{k1}+2^q}_{k,q}, \ ... \ \psi^{x,t_{k2}}_{k,q}] \ , \tag{5}$$

$$\psi^{x,t}_{k,q} = \frac{1}{2^q} \sum_{i=0}^{2^q-1} \psi^{x,t-i}_k \ . \tag{6}$$

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 $[\psi^{x,t}_{k,q}, \ \psi^{y,t}_{k,q}]$ 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.

V. Mezaris, A. Dimou, I. Kompatsiaris, "Local Invariant Feature Tracks for High-Level Video Feature Extraction", Analysis, Retrieval and Delivery of Multimedia Content, N. Adami, A. Cavallaro, R. Leonardi, P. Migliorati (Eds.), Springer Lecture Notes in Electrical Engineering vol. 158, ISBN 978-1-4614-3830-4, 2012. The original publication is available at www.springerlink.com.

Then, the initial trajectory ξ_k can be represented across different time-scales and at various granularity levels as a fixed length vector μ_k:

$$\mu_k = \Big[\lambda(\xi_{k,0}, \frac{\pi}{2}), \quad \lambda(\xi_{k,1}, \frac{\pi}{2}), \ \ldots \lambda(\xi_{k,Q-1}, \frac{\pi}{2})$$
$$\lambda(\xi_{k,0}, \frac{\pi}{4}), \quad \lambda(\xi_{k,1}, \frac{\pi}{4}), \ \ldots \lambda(\xi_{k,Q-1}, \frac{\pi}{4}), \ \ldots$$
$$\lambda(\xi_{k,0}, \frac{\pi}{2^J}), \quad \lambda(\xi_{k,1}, \frac{\pi}{2^J}), \ \ldots \lambda(\xi_{k,Q-1}, \frac{\pi}{2^J})\Big] \ . \tag{7}$$

The corresponding Local Invariant Feature Track (LIFT) descriptor is defined as:

$$LIFT(\psi_k) = [\psi_k^d, \quad \mu_k] \ . \tag{8}$$

The LIFT descriptor is a fixed-length vector that compactly captures both the 2D appearance of a local image region and its long-term motion.

4.2 Invariance Concerns

The definition of the LIFT representation was guided by the need to introduce, to the extent possible, some invariance with respect to the scale and direction of the extracted tracks. Starting with the interest point detection and description in the 2D, the SIFT method was used, due to its well-documented [3, 7] and desirable invariance properties; other similar methods [8, 9] could also be used instead. Concerning the feature track extraction, camera-motion-compensated trajectories were estimated and employed to ensure that the final LIFT representation will not be affected by camera motion. Camera motion could also be useful for representing the shots, but should in any case be separated from the local motion of the different local features within the shot, rather than being allowed to corrupt the latter.

In the subsequent representation of the tracks by histograms, only the direction of each elementary motion of the track was employed, rather than the direction and magnitude of it. This was done for introducing some degree of invariance to image scale, since the same motion (e.g. a person picking up the phone) will result in different motion vector magnitudes depending on the focal length of the camera and its distance from the plane of the motion; on the contrary, the direction of motion is not affected by these parameters.

Histograms at various time-scales were selected for representing the tracks, instead of e.g. comparing the overall displacement of the interest point along the track, to allow for partial matches when considering partial tracks (i.e., when the beginning and end of the different extracted tracks that correspond to the same class of actions do not coincide with each other and with the actual beginning and end of the depicted action). Although the adopted solution may be non-optimal, the reliable matching of partial tracks would otherwise require the use of a computationally expensive optimization-based technique for evaluating the similarity of them, in place of the Euclidean distance typically used in K-Means when creating the "words" used in the Bag-of-Words approach.

The use of motion direction histograms at different granularity levels θ (instead of using a single histogram with a high number of bins) aims at allowing again for partial matches between tracks using a simple metric (i.e., L1/L2 rather than e.g. the Earth Mover's Distance), in the case of small variations in the direction of motion. When considering only a very fine granularity level θ, significant such variations between similar shots could be caused by even small differences in camera angle/viewpoint. The combined use of multiple (from coarse to fine) granularity levels can alleviate this effect to some degree. Alternatively, the weighted assignment of every elementary motion to more than one neighboring bins, when constructing each motion direction histogram, could be employed.

4.3 Shot Representation

The LIFT descriptors of the feature tracks extracted and selected for a video shot, according to the processes of Sect. 3, 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 than simply "visually-similar local regions" (detected by either spatial or spatio-temporal interest point detectors), as in the current

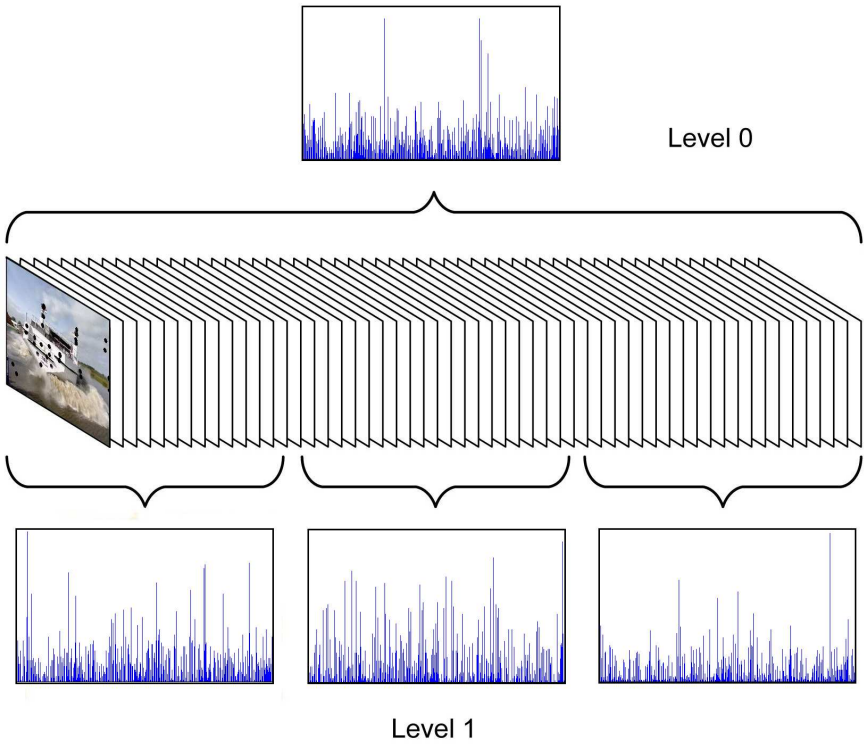

Fig. 4. Illustration of the temporal pyramidal methodology.

state-of-the-art, e.g. [12, 16]. 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. Furthermore, techniques such as spatial pyramids [22] or temporal extensions of them (Fig. 4) can be used in combination with the BoSW model, similarly to the way spatial pyramids are combined with the BoW one.

5 Experimental Results

In the experimental evaluation of the proposed techniques, two datasets were used. The first one is the TRECVID[1] 2007 dataset, which is made of professionally-created videos (Dutch TV documentaries). The training and testing portions of it comprise 50 hours of video each, and 18120 and 18142 shots respectively; all these shots are annotated with 20 concepts that were defined for the TRECVID 2009 contest. This dataset was employed for evaluating different design choices of the proposed BoSW (e.g. the feature track selection strategy) and for comparing them with alternate approaches, as well as for comparing the overall proposed technique with the traditional SIFT-based BoW one. The second dataset is the TRECVID 2010 one, which is made of heterogeneous internet videos. The training and test portions of it comprise approximately 200 hours of video each, and 118536 and 144971 shots respectively; the training portion is annotated with 130 concepts that

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

V. Mezaris, A. Dimou, I. Kompatsiaris, "Local Invariant Feature Tracks for High-Level Video Feature Extraction", Analysis, Retrieval and Delivery of Multimedia Content, N. Adami, A. Cavallaro, R. Leonardi, P. Migliorati (Eds.), Springer Lecture Notes in Electrical Engineering vol. 158, ISBN 978-1-4614-3830-4, 2012. The original publication is available at www.springerlink.com.

were defined for the TRECVID 2010 contest. This dataset was used for further comparing the overall proposed technique with the traditional SIFT-based BoW one, on the basis of the 30 concepts (out of the overall 130 ones) that were evaluated for each run that was submitted to TRECVID 2010.

In the process of extracting the proposed LIFT features of the video shots, the temporal sub-sampling parameter a was set equal to 3. This represents a good compromise between the need for accurately establishing the SIFT point correspondences from frame to frame (which calls for a low value of a, ideally 1) and the need for speeding up the feature extraction process. For each frame of the temporally sub-sampled sequence, the method of [3] was used for interest point detection and description, resulting in a 128-element vector for the local region of each interest point. 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 (7)) for representing the trajectory information of the extracted feature tracks resulted in the LIFT descriptor of each feature track being a 184-element vector, while setting $J = 5$ in selected experiments (indicated below) resulted in a 376-element vector instead.

A first series of experiments was carried out on the TRECVID 2007 dataset, in order 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 Support Vector Machine classifiers. The latter produced a fuzzy class membership degree in the range [0,1] when used for evaluating the relevance of each shot of the TRECVID 2007 test dataset with every one of the considered high-level features, exploiting the BoSW model. Prior to this, the SVM classifiers were trained using the TRECVID 2007 training dataset and the common annotation; for each high-level feature, a single SVM was trained independently of all others. 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 nor in par with SoA works such as [12], 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 (mean average precision, calculated for a maximum of 2000 returned samples per concept [10]) are shown in Fig. 5(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 Sect. 3.2 and of the hypothesis that this process has been based on. Specifically, the selection of the 2500 tracks with the highest probability $p(\psi_k)$, as proposed in Sect. 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 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. For creating the BoW model of each shot, the median frame of the shot was selected as a key-frame and SIFT descriptors were extracted from it. The results (Fig. 5(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 than the latter when considering only dynamic concepts (i.e., a subset of the 20 defined high-level features, which is discussed in more detail below).

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 [12]. 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, the temporal pyramid of Fig. 4 for BoSW. Additionally, in the latter case 5 granularity levels θ (i.e., $J = 5$ in (7)), 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. Overall, considering the second of the the two tested configurations (500 words, soft assignment, spatial/temporal pyramidal decomposition), the SIFT-based BoW resulted in a mean average precision (MAP) of 0.084, whereas the combination of BoW and

V. Mezaris, A. Dimou, I. Kompatsiaris, "Local Invariant Feature Tracks for High-Level Video Feature Extraction", Analysis, Retrieval and Delivery of Multimedia Content, N. Adami, A. Cavallaro, R. Leonardi, P. Migliorati (Eds.), Springer Lecture Notes in Electrical Engineering vol. 158, ISBN 978-1-4614-3830-4, 2012. The original publication is available at www.springerlink.com.

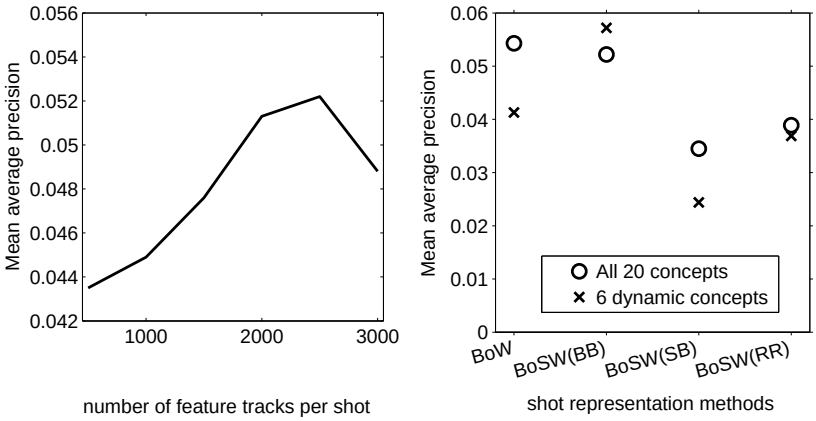

Fig. 5. Evaluation of a) the impact of the number of feature tracks used for representing each shot, and b) the impact of different shot representation techniques, on concept detection performance.

BoSW in a MAP of 0.102, representing an increase of the former by approximately 21%. Considering only high-level features that have a strong temporal dimension ("people-dancing", "person-playing-soccer", etc.), i.e. features 5, 6, 7, 9, 11 and 13 of Fig. 6, the use of the proposed BoW and BoSW combination leads to an increase of MAP by approximately 28% over using the SIFT-based BoW alone. The significance of taking into account motion information, as done by BoSW, for detecting such dynamic concepts can also be seen in Fig. 6, where the per-concept results (average precision) corresponding to the last row of Table 1 are shown.

Table 1. Comparison between BoW, combination of BoW and BoSW on the TRECVID 2007 dataset (mean average precision for all 20 / 6 dynamic concepts).

	BoW		BoW+BoSW(BB)	
Number of considered concepts:	20	6	20	6
500 words, hard assignment	0.054	0.041	0.068	0.056
500 words, soft assignment, spatial/temporal pyramidal decomposition	0.084	0.088	0.102	0.113

Finally, the SIFT-based BoW and the combination of BoW and BoSW (using again 500 words, soft assignment, spatial/temporal pyramidal decomposition, and 5 granularity levels θ) were compared on the TRECVID 2010 dataset, by participating with the two corresponding runs to the TRECVID 2010 contest [23]. The results for the 30 concepts that were evaluated in this contest are reported in Table 2 and Fig. 7 (overall and per-concept results, respectively). Extended inferred average precision (xinfAP) and mean extended inferred average precision (MxinfAP) [24], calculated for a maximum of 2000 returned samples per concept, were used for quantifying the results, in order to account for the test portion of this dataset being annotated only in part. It can be seen that the SIFT-based BoW resulted in a MinfAP of 0.030, whereas the combination of BoW and BoSW in a MinfAP of 0.038, representing an increase of the former by approximately 26.7%. Considering only high-level features that have a strong temporal dimension, i.e. features 1, 4, 7, 11, 23, 26, 28, and 30 of Fig. 7, the use of the proposed BoW and BoSW combination leads to an increase of MinfAP by approximately 95% over using the SIFT-based BoW alone.

V. Mezaris, A. Dimou, I. Kompatsiaris, "Local Invariant Feature Tracks for High-Level Video Feature Extraction", Analysis, Retrieval and Delivery of Multimedia Content, N. Adami, A. Cavallaro, R. Leonardi, P. Migliorati (Eds.), Springer Lecture Notes in Electrical Engineering vol. 158, ISBN 978-1-4614-3830-4, 2012. The original publication is available at www.springerlink.com.

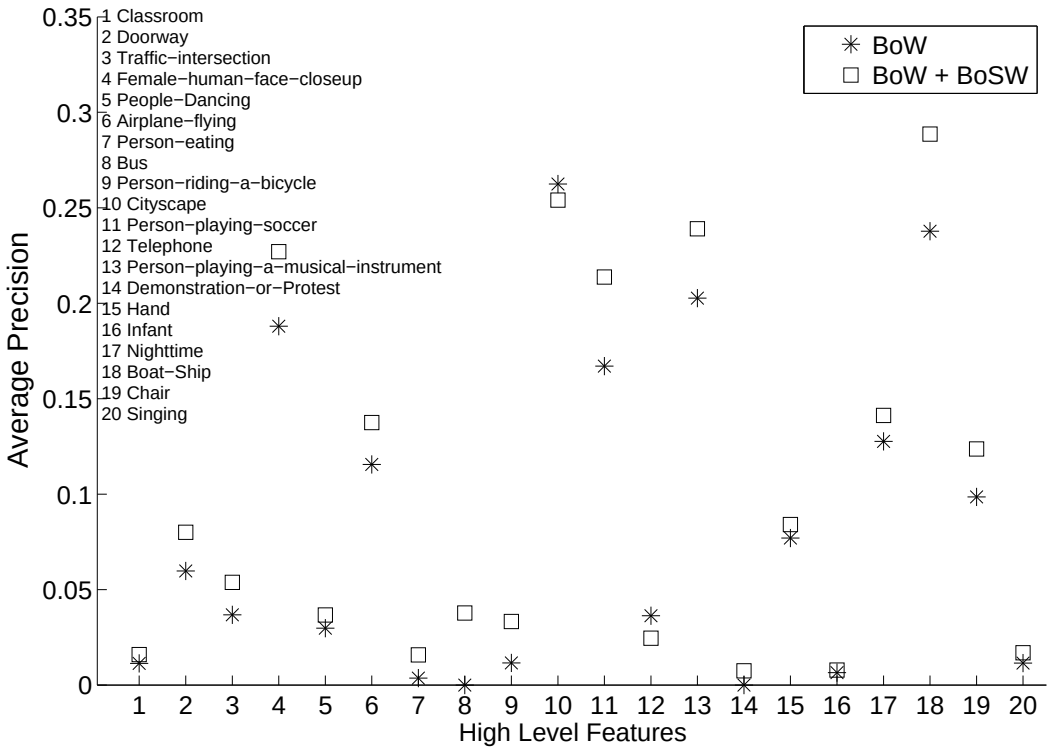

Fig. 6. Individual concept detection results on the TRECVID 2007 dataset for BoW alone and for the combination of BoW and BoSW, using 500 words, soft assignment, and spatial/temporal pyramidal decomposition.

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. In particular, techniques for the extraction, selection, representation and use of feature tracks for the purpose of constructing a Bag-of-Spatiotemporal-Words model for the video shots were presented. Experimental evaluation of the proposed approach on two challenging test corpora (TRECVID 2007, TRECVID 2010) revealed its potential for concept detection in video, particularly when considering dynamic rather than static concepts.

Acknowledgements

This work was supported by the European Commission under contract FP7-248984 GLOCAL.

References

1. Mezaris, V., Kompatsiaris, I., Boulgouris, N., Strintzis, M.: Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval. IEEE Trans. on Circuits and Systems for Video Technology **14**(5) (May 2004) 606–621
2. Mezaris, V., Kompatsiaris, I., Strintzis, M.: Video Object Segmentation using Bayes-based Temporal Tracking and Trajectory-based Region Merging. IEEE Trans. on Circuits and Systems for Video Technology **14**(6) (June 2004) 782–795
3. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. of Computer Vision **60** (2004) 91–110

V. Mezaris, A. Dimou, I. Kompatsiaris, "Local Invariant Feature Tracks for High-Level Video Feature Extraction", Analysis, Retrieval and Delivery of Multimedia Content, N. Adami, A. Cavallaro, R. Leonardi, P. Migliorati (Eds.), Springer Lecture Notes in Electrical Engineering vol. 158, ISBN 978-1-4614-3830-4, 2012. The original publication is available at www.springerlink.com.

Table 2. Comparison between BoW, combination of BoW and BoSW on the TRECVID 2010 dataset (mean extended inferred average precision for all 30 / 8 dynamic concepts).

	BoW		BoW+BoSW(BB)	
Number of considered concepts:	30	8	30	8
500 words, soft assignment, spatial/temporal pyramidal decomposition	0.030	0.020	0.038	0.039

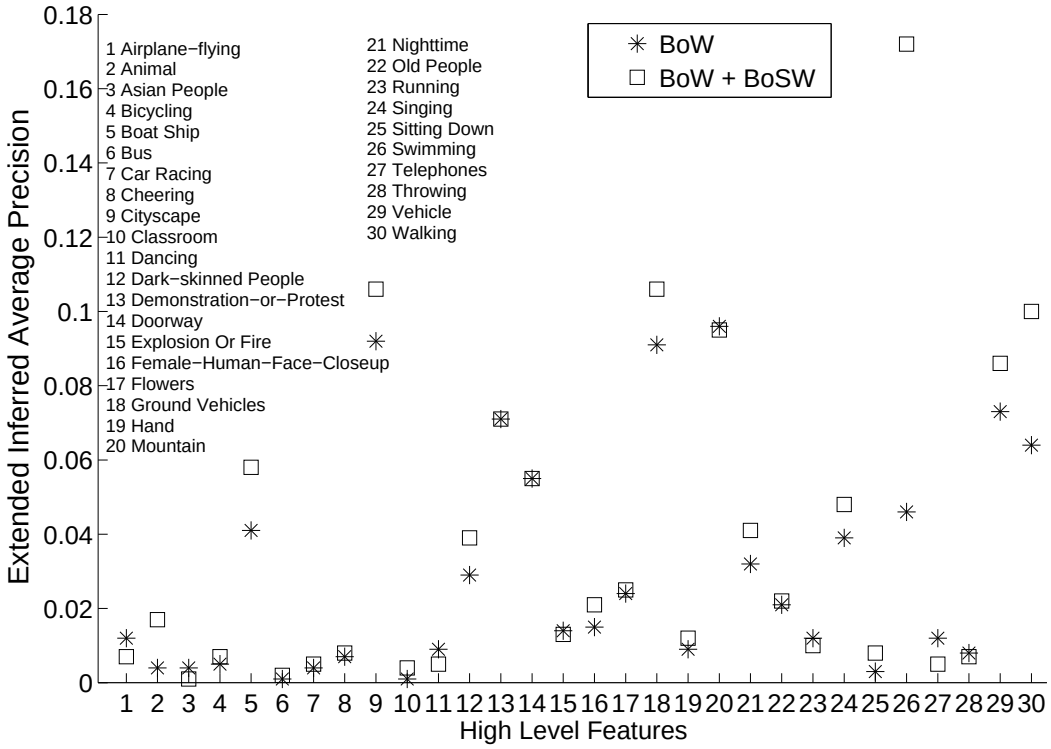

Fig. 7. Individual concept detection results on the TRECVID 2010 dataset for BoW alone and for the combination of BoW and BoSW, using 500 words, soft assignment, and spatial/temporal pyramidal decomposition.

4. Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: Proc. ECCV Int. Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic (May 2004)
5. Mezaris, V., Sidiropoulos, P., Dimou, A., Kompatsiaris, I.: On the use of visual soft semantics for video temporal decomposition to scenes. In: Proc. Fourth IEEE Int. Conf. on Semantic Computing (ICSC 2010), Pittsburgh, PA, USA (September 2010)
6. Gkalelis, N., Mezaris, V., Kompatsiaris, I.: Automatic event-based indexing of multimedia content using a joint content-event model. In: Proc. ACM Multimedia 2010, Events in MultiMedia Workshop (EiMM10), Firenze, Italy (October 2010)
7. Mikolajczyk, K., Schmid, C.: A Performance Evaluation of Local Descriptors. IEEE Trans. on Pattern Analysis and Machine Intelligence **27**(10) (2005) 1615–1630
8. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. Computer Vision and Image Understanding **110**(3) (2008) 346–359
9. Burghouts, G.J., Geusebroek, J.M.: Performance Evaluation of Local Colour Invariants. Computer Vision and Image Understanding **113** (2009) 48–62
10. Smeaton, A.F., Over, P., Kraaij, W.: High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In Divakaran, A., ed.: Multimedia Content Analysis, Signals and Communication Technology.

Springer (2009) 151–174
11. Piro, P., Anthoine, S., Debreuve, E., Barlaud, M.: Combining spatial and temporal patches for scalable video indexing. Multimedia Tools and Applications **48**(1) (2010) 89–104
12. Snoek, C., van de Sande, K., de Rooij, O., et.al.: The MediaMill TRECVID 2008 Semantic Video Search Engine. In: Proc. TRECVID 2008 Workshop, USA (Nov. 2008)
13. Ballan, L., Bertini, M., Bimbo, A.D., Serra, G.: Video Event Classification using String Kernels. Multimedia Tools and Applications **48**(1) (2010) 69–87
14. Chen, M., Hauptmann, A.: MoSIFT: Recognizing human actions in surveillance videos. Technical Report CMU-CS-09-161, Carnegie Mellon University (2009)
15. Laptev, I.: On Space-Time Interest Points. Int. J. of Computer Vision **64**(2/3) (2005) 107–123
16. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. Int. J. of Computer Vision **79**(3) (Sept. 2008) 299–318
17. Zhou, H., Yuan, Y., Shi, C.: Object tracking using SIFT features and mean shift. Computer Vision and Image Understanding **113**(3) (2009) 345–352
18. Tsuduki, Y., Fujiyoshi, H.: 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)
19. Anjulan, A., Canagarajah, N.: A Unified Framework for Object Retrieval and Mining. IEEE Trans. on Circuits and Systems for Video Technology **19**(1) (Jan. 2009) 63–76
20. Moenne-Loccoz, N., Bruno, E., Marchand-Maillet, S.: Local Feature Trajectories for Efficient Event-Based Indexing of Video Sequences. In: Proc. Int. Conf. on Image and Video Retrieval (CIVR), Tempe, USA (July 2006)
21. Sun, J., Wu, X., Yan, S., Cheong, L., Chua, T.S., Li, J.: Hierarchical Spatio-Temporal Context Modeling for Action Recognition. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR), Miami, USA (June 2009)
22. Lazebnik, S., Schmid, C., Ponce, J.: Spatial pyramid matching. In Dickinson, S., Leonardis, A., Schiele, B., Tarr, M., eds.: Object Categorization: Computer and Human Vision Perspectives. Cambridge University Press (2009)
23. Moumtzidou, A., Dimou, A., Gkalelis, N., Vrochidis, S., Mezaris, V., Kompatsiaris, I.: ITI-CERTH participation to TRECVID 2010. In: Proc. TRECVID 2010 Workshop, USA (Nov. 2010)
24. Yilmaz, E., Kanoulas, E., Aslam, J.: A simple and efficient sampling method for estimating AP and NDCG. In: Proc. 31st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR). (2008) 603–610