IEEE International Conference on Image Processing (ICIP 2013), Melbourne, Australia, September 2013

ENHANCING VIDEO CONCEPT DETECTION WITH THE USE OF TOMOGRAPHS

Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris

Information Technologies Institute / Centre for Research and Technology Hellas
6th Km Charilaou-Thermi Road, P.O.BOX 60361, Thermi 57001, Greece

ABSTRACT

In this work we deal with the problem of video concept detection, for the purpose of using the detection results towards more effective concept-based video retrieval. In order to handle this task, we propose using spatio-temporal video slices, called video tomographs, in the same way that visual keyframes are typically used in traditional keyframe-based video concept detection schemes. Video tomographs capture in a compact way motion patterns that are present in the video, and are used in this work for training a number of base detectors. The latter augment the set of keyframe-based base detectors that can be trained on different image representations. Combining the keyframe-based and tomograph-based detectors, improved concept detection accuracy can be achieved. The proposed approach is evaluated on a dataset that is extensive both in terms of video duration and concept variation. The experimental results manifest the merit of the proposed approach.

Index Terms— video analysis, supervised learning, support vector machines, video tomograph, concept detection

1. INTRODUCTION

One of the main goals of the image and video processing community is to develop techniques that would allow the automatic understanding of unconstrained video. In the root of this task lies the fast and accurate detection of the semantic concepts depicted in the video. The efficient and effective detection of concepts by looking purely at the visual content is an important and challenging problem.

In the last years, the research community, partially triggered by the TRECVID Semantic Indexing task [1], has been shifting its focus on large-scale video concept detection, i.e. the development of systems that would be able to handle large amounts of video data and detect multiple semantic concepts efficiently (e.g. [2], [3]). As a result, several powerful techniques have emerged, which aim to combine high precision and low computational cost. For example, in order to exploit color information in addition to local image structure, the Opponent-SIFT and RGB-SIFT (or Color-SIFT) variations of the well-known SIFT descriptor [4] were proposed in [5]. Furthermore, in order to reduce computational cost, Speeded Up Robust Features (SURF) [6] and DAISY [7] were introduced as fast SIFT approximations; interest point detection (traditionally performed with the help of corner detectors, e.g. the Harris-Laplace one [8]) was fully or partially replaced in many schemes by dense sampling (i.e. the sampling of image patches on a regular dense grid); and chi-square kernels, that were originally considered to be optimal for use in support vector machines (SVM) [9], [10] are now often replaced by Histogram Intersection kernels [11] or even Linear SVMs, to name a few recent developments in this area.

Contrary to what is intuitively expected, in most of the developed schemes that aim to detect multiple concepts in video data, motion information is ignored and the detection is based exclusively on a set of characteristic keyframes that are extracted at shot level (i.e. each video shot is represented by one or more keyframes). This is explained by the fact that motion descriptor extraction is typically associated with high computation cost, and the gains in precision that are attained by introducing motion descriptors in the concept detection process are often disproportionally low, compared to the added computational complexity. However, a concept detection algorithm that uses no motion information handles the video stream as a mere collection of photos (keyframes), failing to take advantage of the dynamic nature of video that makes it particularly expressive.

In this work we propose the use of video tomographs [12] (i.e. spatio-temporal slices with one axis in time and one in space) to represent video motion patterns. These tomographs are straightforwardly extracted, their extraction exhibits extremely low computation cost, and as we show in this work they can then be analyzed as if they were regular keyframes. We demonstrate that video tomographs, when used along with visual keyframes, enhance video concept detection while being a computationally efficient solution towards exploiting information about the temporal evolution of the video signal.

The rest of the paper is organized as follows. An overview of the proposed approach is presented in Section 2. The use of visual tomographs for concept detection is detailed in Section 3. The experimental results are reported in Section 4 and, finally, conclusions are drawn in Section 5.

2. OVERVIEW OF THE PROPOSED APPROACH

The pipeline of a typical concept detection system is shown in Fig. 1. The video stream is initially sampled, for instance by selecting one or multiple keyframes per shot. Subsequently, each sample is represented using one or more types of appropriate features (e.g. SIFT [4], SURF [6]). These features form the input to a number of base classifiers, which typically rely on vector quantization and support vector machines. The parameter sets that control the employed classifiers are predefined (i.e. have been learned at the classifier training stage), using similar features extracted from training data. Finally, the base classifier outputs are fused to estimate a final concept detection score. It should be noted that if multiple concepts are to be detected, this process is executed multiple times, independently for each of the considered concepts.

In this work we focus on the first component of the analysis pipeline, i.e. the video sampling, to extract not only keyframes but also video tomographs. We then show that all other components of the analysis pipeline that processes the extracted video samples (both keyframes and tomographs) can follow established state-of-the-art approaches. More specifically, in our work the employed rep-

This work was supported by the European Commission under contracts FP7-287911 LinkedTV, FP7-600826 ForgetIT and FP7-318101 MediaMixer.

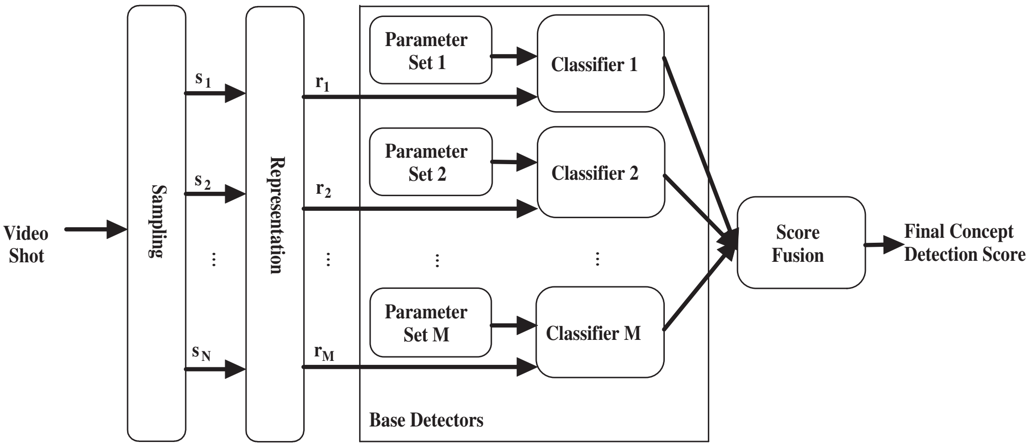

Fig. 1. The pipeline of a typical concept detection system. Initially the video stream is sampled (e.g. keyframes are extracted) using N different sampling strategies (labeled s_1, s_2,... s_N in the figure). Subsequently, M sets of features are extracted to represent the visual information samples (labeled r_1, r_2,...r_M in the figure). The set of features are used as inputs to base classifiers that are trained off-line. Finally, the base classifier outputs are combined and an overall concept detection score is estimated.

resentations are SIFT, RGB-SIFT and Opponent-SIFT, which were experimentally found in [5] to form the optimal low-level visual descriptor set for video concept detections tasks. These descriptors are extracted from local image patches. Similarly to the current state of the art, we use two approaches for selecting these patches. In the first one the interest points are selected through dense sampling, while in the second one interest point detection is performed through a Harris-Laplace corner detector [8]. The extracted low-level descriptors are assigned to visual words using separately two vocabularies, which were created off-line by k-means clustering, through hard-assignment and soft-assignment, respectively [13]. A pyramidal 3 × 1 decomposition scheme, employing 3 equally-sized horizontal bands of the image [14], is used in all cases, thus generating 3 different Bag-of-Words (BoW) feature vectors from image bands, while a fourth BoW is built using the entire image. In all cases, the number of words for each BoW was set to 1000. Thus, for each combination of video sampling strategy, interest point detector, descriptor and assignment method a vector of 4000 dimensions is finally extracted and used as the actual input to the utilized base classifiers. The latter are linear SVMs, chosen so as to keep low the required computations time. All classifiers were trained off-line, using the extensive training data that were made available as part of the TRECVID 2012 Semantic Indexing task [15].

3. VIDEO TOMOGRAPHS FOR CONCEPT DETECTION

In this section we explain how keyframe-based concept detection can be improved by augmenting the set of keyframes with a spatio-temporal type of image, the video tomograph. Video tomographs were introduced in [12] as spatio-temporal slices and have been used for optical flow estimation [16], camera motion classification [17] and video copy detection [18], [19]. A video tomograph is defined in [12] as a cross-section image (i.e. an image defined by the intersection between a plane and the video volume) which is additionally smoothed using a high-pass filter. The cross-section image is generated by fixing a 1-D line on the image plane and aggregating the video content falling on the corresponding line for all frames of the shot. In this work video tomographs are re-defined in a slightly different and somewhat more general way, and are used in a completely new way for a different application.

Video tomograph re-definition is based on the fact that the video volume is not continuous, but is formed by a finite set of frames. Consequently, tomographs can be defined as a set of line segments, which are recursively estimated as intersections between lines and frames. More specifically, if f_i is the current frame, v_{i-1} the line defining the intersection in the previous frame, R_i the current tomograph rotation matrix and T_i the current tomograph translation vector then the $i-th$ line is estimated as:

$$v_i = f_i \cap (R_i * v_{i-1} + T_i) \tag{1}$$

If v_0 is the initial line segment and all R_i, T_i are known, then a tomograph image can be straightforwardly extracted. It should be noted that this tomograph definition encompasses the definition of [12]. As a matter of fact, this corresponds to the selection $R_i = I_2, T_i = [0\ 0]^T\ \forall i$, where I_2 is the two-dimensional identity matrix and the superscript T denotes the transpose matrix.

Based on the above definition, complex motion patterns can be projected into meaningful images. For example, a tomograph could be formed by lines chosen so as to be always perpendicular to the camera motion direction, thus generating an image that captures the objects being followed by the camera. While such an approach would require the characterization of camera motion, several methods exist for automatically detecting camera motion parameters from the video (e.g. [20], [21]), and such methods have also been used directly on the motion vectors encoded in the MPEG stream (e.g. [22]), thus introducing minimal computational overhead to the overall video analysis pipeline.

Putting aside for now the possibility of taking into account camera motion, the two most simple tomograph images are the centralized horizontal (CH-tomograph) and the centralized vertical (CV-tomograph) tomographs. A CH-tomograph is constructed by aggregating for all frames of a shot the visual content of the horizontal line passing from the frame center (i.e. $R_i = I_2, T_i = [0\ 0]^T\ \forall i$ and v_0 is the line $y = H/2$, where H is the frame height). A CV-tomograph is constructed in an analogous way, with the only difference being that v_0 is perpendicular to the x-axis, instead of parallel to it.

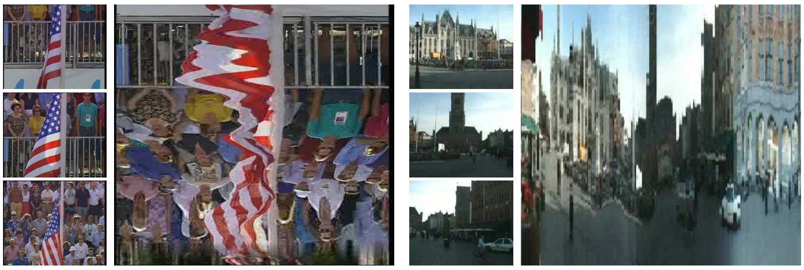

Fig. 2. Two tomograph examples, each one corresponding to a different type of tomograph image. The left tomograph is a CH-tomograph, while the right a CV-tomograph. Both of them are defined by the temporal ordering of lines that pass from the center of the frame. Three indicative frames of the shot from which each tomograph was generated are also shown to the left of the corresponding tomograph (the temporal order of the shown frames is from the top to the bottom).

In Fig. 2 a CH-tomograph and a CV-tomograph example are shown. In the left example the shot shows the national anthem ceremony in a sports event. As the camera follows the raising flag, the CH-tomograph “draws” a flipped version of the scene background. The flipping artifact is not expected to play an important role in the following steps of the concept detection algorithm, since most of the well-known low-level descriptors are orientation invariant. On the other hand, in the right example the video shot depicts a road junction. In this case, the camera is moving in the horizontal direction. The CV-tomograph, which is generated by lines perpendicular to the camera motion direction, generates a “mosaic-like” image of the urban scene.

For the purpose of concept detection, the tomographs are processed in the same way as keyframes. More specifically, image patches are estimated, followed by descriptor extraction and vector quantization. It should be noted that the vocabulary employed at this stage is constructed by clustering visual words extracted from the corresponding tomograph type (e.g. a random sample of CV-tomograph SIFT vectors are clustered in order to generate the vocabulary used for vector quantization of SIFT descriptors extracted from CV-tomograph images). The resulting Bag-of-Words feature vectors are the input to tomograph-based base classifiers. These classifiers are also independently trained for each tomograph type, using annotated samples taken from tomographs of the corresponding type. Finally, the base classifier output is fused with the output of the keyframe-based classifiers in a simple averaging scheme that does not discriminate between outputs of keyframe and tomograph-based classifiers.

Concerning the computational cost of introducing video tomographs in the concept detection process, it is straightforward that this processing time depends on the total number of pixels in each tomograph. Consequently, an estimation of the tomograph size can be used to compare the computational cost of tomograph-based classification with the computational cost of keyframe-based classification. Keyframe size is constant for a given video and can be adjusted during the decoding process. On the other hand, tomograph size is not constant, since it depends not only on frame size and frame ratio (that are typically constant) but also on the current shot duration. However, a rough estimation of the mean tomograph size is possible, at least for CH-tomographs and CV-tomographs. As a matter of fact, if W and H is the frame width and height, r is the frame rate and τ_s the duration of shot s then the total number of pixels for keyframe K, CH-tomograph K_H and CV-tomograph K_V would be:

$$\#(K) = WH \qquad (2)$$

$$\#(K_H) = \lfloor r\tau_s \rfloor W, \quad \#(K_V) = \lfloor r\tau_s \rfloor H \qquad (3)$$

where $\#(.)$ operator denotes the number of frames and $\lfloor . \rfloor$ the integer part of a real number. In the extensive TRECVID SIN 2012 dataset, the mean shot duration is 5.1 seconds. If typical values ($r = 25$, $\tau_s = 5.1$, $W = 352$, $H = 288$) are replaced in the above equations then the number of pixels of both CH-tomographs and CV-tomographs compared to the number of pixels in a keyframe would be:

$$(\#(K_V) + \#(K_H))/\#(K) \simeq 0.8 \qquad (4)$$

Consequently, the descriptor extraction computational cost when using a pair of tomographs is similar to the cost of processing a single keyframe. This cost is minimal compared to typical motion descriptors, since the extraction of the latter involves processing all frames of each shot or a large subset of them. Due to this, the extraction of even computationally efficient spatio-temporal descriptors (e.g. [23], [24]) is much more computationally demanding.

Finally, it should be noted that the creation of tomographs is also very fast. Apart from decoding the video stream into frames, this requires only accessing a small set of image pixels. This set is determined for each frame by Eq. (1) in $O(max\{W, H\})$ cost.

4. EVALUATION AND EXPERIMENTAL RESULTS

To examine the contribution of tomographs towards more accurate concept detection, we conducted an experimental comparison of a concept detection scheme that employs only 1 keyframe per shot and a concept detection scheme that additionally employs 1 CH-tomograph and 1 CV-tomograph per shot. We selected these two simple tomographs for our experiments in order to demonstrate that tomographs can enhance performance even if a non-optimized, simple tomograph extraction method is followed. Additionally, a third configuration in which only the aforementioned tomographs are used was also included in the comparison. In all cases, the scores of the different base classifiers (regardless of whether these are keyframe-based classifiers, tomograph-based ones, or a mixture of both types) were fused at the last stage of the concept detection process simply by calculating their harmonic mean.

The experimental setup employs the entire video dataset and concept set that were used in the 2012 TRECVID SIN task. More specifically, 46 semantic concepts were evaluated. The detection of these concepts takes place in a video dataset comprising 8263

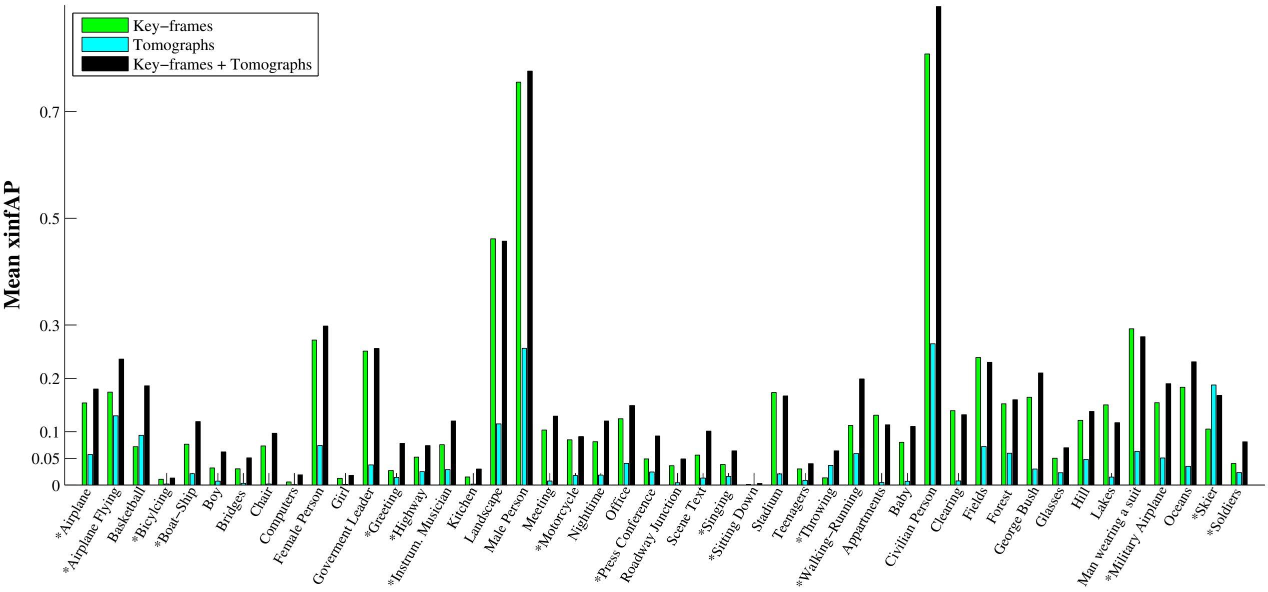

Fig. 3. Performance comparison of a concept detection system that uses tomographs plus keyframes versus a system that uses exclusively keyframes, or exclusively tomographs, in TRECVID 2012 Semantic Indexing dataset. Concept detection accuracy is measured by xinfAP.

videos of almost 200 hours total duration. The whole dataset is offline pre-segmented into more than 140 thousand shots. The goal of each concept detector is to retrieve the top-2000 shots that are most likely for the concept to be present. The 2000 shots are sorted using the detectors' score in descending order and the results are evaluated using partial, manually generated ground-truth annotations. The employed detection accuracy measure is Extended Inferred Average Precision (xinfAP) [25], which is a measure approximating Average Precision, when the ground-truth annotations are not complete. The employed ground-truth annotations and the xinfAP implementation are the ones provided by the TRECVID organizers.

The experimental results are shown for each concept in Fig. 3. Although many of the 46 concepts are not intuitively expected to be strongly correlated with any type of motion (e.g. "landscape", "fields", "computers") we can see from this figure that combining keyframe- and tomograph-based concept detection increases the accuracy for 39 of the 46 concepts. Overall, the performance as measured by mean xinfAP increases from 0.135 to 0.156, representing a 15.5% accuracy boost. This together with the standalone performance of video tomographs, which is expressed by a mean xinfAP of 0.044, show that although the tomographs are not potential replacements of the keyframes, they provide additional information that the latter do not capture, thus being a valuable addition to keyframe-based concept detection approaches.

Furthermore, these results indicate that using tomographs in addition to one or a few keyframes is beneficial, compared to using a large number of keyframes for each shot. In [3], a concept detection scheme similar to our baseline keyframe-based approach was employed, in two versions differing only in that the first one exploited only 1 keyframe for each shot, while the second employed 10 additional keyframes. The accuracy boost achieved by the second version in relation to the first one was 14.7%, which is comparable to the one achieved in our work by the introduction of a pair of tomographs, but the associated computational cost of using an extra 10 keyframes per shot is higher than the cost of using a pair of tomographs by one order of magnitude.

Finally, it should be noted that the concepts that benefit the most from the introduction of tomographs are, as expected, the dynamic concepts, i.e. those that are clearly related with motion. In the employed concept set we have identified 15 concepts that are either directly related with actions that involve motion (e.g. "throwing", "walking-running", "bicycling") or are objects that are very likely to be filmed while they are in motion (e.g. "skier", "motorcycle", "boat-ship"). In Fig. 3 these concepts are marked with a "*". If only these concepts are taken into account, the accuracy boost caused by introducing tomographs is 56.4% (mean xinfAP rising from 0.074 to 0.116). For the remaining, rather static concepts, the corresponding mean xinfAP boost is limited to 11%. Among the latter concepts, there are 7 for which the introduction of tomographs results in detection accuracy reduction (i.e. "landscape", "stadium", "apartments", "clearing", "fields", "lakes", "man wearing a suit"). The reason for this is that shots associated with such concepts typically do not exhibit specific motion patterns that could contribute to the detection of the concept. Consequently, for these concepts the tomograph-based classifiers primarily introduce noise to the overall detection pipeline.

5. CONCLUSIONS

In this work we examined the use of video tomographs as an additional sampling strategy for the video concept detection task. Our experimental results give evidence that the use of video tomographs can be an efficient and effective way to include motion-related information in concept detection systems, while it is expected that introducing more sophisticated tomograph selection techniques in the future could further increase the achieved accuracy.

6. REFERENCES

[1] 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, Ajay Divakaran, Ed., pp. 151–174. Springer Verlag, Berlin, 2009.

[2] A. Wei, Y. Pei, and H. Zha, "Random-sampling-based spatial-temporal feature for consumer video concept classification," in *Proc. of the 2012 IEEE International Conference on Image Processing (ICIP)*, 2012, pp. 1861–1864.

[3] C.G.M. Snoek, K.E.A. Sande, X. Li, M. Mazloom, Y.-G. Jiang, D.C. Koelma, and A.W.M. Smeulders, "The MediaMill TRECVID 2011 semantic video search engine," in *Proceedings of the 9th TRECVID Workshop*, Gaithersburg, USA, December 2011.

[4] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] K.E.A. Sande, T. Gevers, and C.G.M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[6] H. Bay, T. Tuytelaars, and L.V. Gool, "Surf: Speeded up robust features," in *Proc. of European Conference on Computer Vision*, 2006, pp. 404–417.

[7] E. Tola, V. Lepetit, and P. Fua, "A fast local descriptor for dense matching," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[8] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. of 4th Alvey Vision Conference*, 1988, pp. 147–151.

[9] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 213–238, 2007.

[10] Y.G. Jiang, C.W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. of the 6th ACM International Conference on Image and Video Retrieval*, 2007, pp. 494–501.

[11] S. Maji, A.C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[12] Y. Tonomura and A. Akutsu, "Video tomography: An efficient method for camerawork extraction and motion analysis," in *Proc. of Second ACM International Conference on Multimedia (ACM MM 1994)*, 1994, pp. 349–356.

[13] J. Gemert, C.J. Veenman, A. Smeulders, and J. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[14] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 2169–2178.

[15] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A.F. Smeaton, and G. Queenot, "Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2012*, 2012.

[16] A. Hauptmann and M. Smith, "Text, speech, and vision for video segmentation: The informedia project," in *AAAI Fall Symposium, Computational Models for Integrating Language and Vision*, 1995.

[17] W. Jiang and A. Loui, "Video concept detection by audio-visual grouplets," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 4, pp. 223–238, 2012.

[18] G. Leon, H. Kalva, and B. Furht, "Video identification using video tomography," in *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, 2009, pp. 1030–1033.

[19] H.-S. Min, S. Kim, W.D. Neve, and Y.M. Ro, "Video copy detection using inclined video tomography and bag-of-visual-words," in *Proc. of the 2012 IEEE International Conference on Multimedia and Expo (ICME)*, 2012, pp. 562–567.

[20] T. Yu and Y. Zhang, "Retrieval of video clips using global motion information," *Electronics Letters*, vol. 37, no. 14, pp. 893–895, 2001.

[21] G.B. Rath and A. Makur, "Iterative least squares and compression based estimations for a four-parameter linear global motion model and global motion compensation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 7, pp. 1075–1099, 1999.

[22] 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 Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 606–621, 2004.

[23] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. of British Machine Vision Conference (BMVC)*, 2009.

[24] G. Willems, T. Tuytelaars, and L.V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. of the 10th European Conference on Computer Vision (ECCV): Part II*, 2008, pp. 650–663.

[25] E. Yilmaz, E. Kanoulas, and J.A. Aslam, "A simple and efficient sampling method for estimating ap and ndcg," in *Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008, pp. 603–610.