Compact Video Description and Representation for Automated Summarization of Human Activities

Ioannis Mademlis, Anastasios Tefas, Nikos Nikolaidis and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

***Abstract*—A compact framework is presented for the description and representation of videos depicting human activities, with the goal of enabling automated large-volume video summarization for semantically meaningful key-frame extraction. The framework is structured around the concept of per-frame visual word histograms, using the popular Bag-of-Features approach, and the spatial pyramid image partitioning scheme. Three existing image descriptors (histogram, FMoD, SURF) and a novel one (LMoD), as well as a component of an existing state-of-the-art activity descriptor (Dense Trajectories), are adapted into the proposed framework and quantitatively compared against each other, as well as against the most common video summarization descriptor (global image histogram), using a publicly available annotated dataset and the most prevalent video summarization method, i.e., video frame clustering. In all cases, several image channels are exploited (luminance, hue, edges, optical flow magnitude) in order to simultaneously capture information about the depicted shapes, colors, lighting, textures and motions. The quantitative evaluation results indicate that one of the proposed descriptors clearly outperforms the competing approaches in the context of the presented framework.**

Keywords*—*Video Summarization, Video Description, Key-Frame Extraction

I. Introduction

Several applications exist nowadays where large-scale video footage depicting human activities needs to be analyzed, possibly on a frame-by-frame basis, requiring human intervention. Examples include professional capture sessions, where the action described in the script is typically filmed using multiple cameras, or streams from surveillance cameras which may be capturing continuously for many days. A very large volume of data are usually produced in such scenarios, which may well exceed 6TB per day [1]. This amount of data is difficult to be efficiently assessed and analysed manually, demanding a great deal of human effort.

Video summarization can be employed as an automated solution to such problems, by generating a condensed version of the video that only contains the most important content [2]. Subsequently, these summaries may be used instead of the original video streams in order to alleviate storage or computational requirements, or the necessary human labour, e.g., in case manual annotation is needed. Most summarization methods initially select a subset of important video frames (*key-frames*) that compactly represent the entire video content. The abstracted content that needs to be included in the target summary can be represented either as an ordered set of static key-frames, or as a dynamic video skim, with the former being more suitable for indexing, browsing and retrieval applications [3]. Evaluation of the success of a summarization method is typically subjective, due to the inherently subjective nature of the task.

Video frames are initially described by low-level image descriptors, such as global color-based, texture-based or shape-based features [4]. In general, the most commonly employed video frame descriptors are variants of global joint image histograms in the HSV color space [5] [6]. In order to bring down the computational requirements of the subsequent summarization process, dimensionality reduction on such color histograms has been attempted [7]. In [8] the low-level Frame Moments Descriptor (FMoD) is introduced, a video descriptor designed for compactly capturing statistical characteristics of several image channels, both in a global and in various local scales. In a number of works [3] [9], local, recognition-oriented image descriptors have been employed for video description (e.g., Scale-Invariant Feature Transform (SIFT) [10] or Speeded-Up Robust Features (SURF) [11]), using the popular Bag-of-Features (BoF) representation model [12]. In general, video description and representation methods specifically suited to video summarization have not been studied extensively.

Key-frame extraction typically includes clustering the video frame descriptors into groups. Subsequently, a set of frames that are closest to each cluster centroid are initially selected as key-frames. In many cases, information about the way a video is naturally segmented into shots (e.g., in movies [8]) is also exploited to assist the summarization process [13] [6], e.g. by applying clustering at shot-level. Typically, a number of the extracted key-frames are filtered out to reduce redundancy and the rest are presented in temporal order.

Although the above approaches are oriented towards generic video input, methods exploiting video type-specific information have also been proposed. In surveillance videos, temporal segmentation (shot boundaries detection [14]) is not a viable option due to the lack of cuts, therefore motion detection is employed in order to create summaries that contain sets of object actions, like pedestrian walking. Detected actions taking place in different direction and speed, are fused into a single scene to form a short length video or graphical cue containing as many actions as possible [15]. However, in unedited videos from professional capture sessions (e.g., in TV/movie production), which are also filmed with a static camera and not clearly segmented into shots, such an approach is not applicable. The preferred summarization goal would naturally be to select one key-frame per depicted activity. Moreover, while in both cases the most important visual clue is human motion, state-of-the-art human activity descriptors such as Dense Trajectories [16] cannot be readily employed, due

to the exceptionally high memory/computational requirements and their unsuitability for per-frame descriptions, given that accurate activity descriptions extend temporally in multiple neighbouring video frames.

This work attempts to investigate video description/representation specifically suited to video summarization tasks. It presents a compact framework for the description and representation of videos depicting human activities, with the goal of enabling automated large-volume video summarization for semantically meaningful key-frame extraction. The goal is to succinctly summarize a video by, ideally, selecting one key-frame per depicted activity segment. The framework is structured around the concept of per-frame visual word histograms, using the established BoF approach. This scheme successfully captures the distribution of elemental visual building blocks at each video frame and has been proven suitable for discriminative representation of human activities, in tasks such as activity recognition [17] or temporal activity segmentation [18].

Three existing image descriptors (histogram, FMoD, SURF), a novel one (LMoD), as well as a component of Dense Trajectories, are adapted into the proposed framework. The image descriptors are quantitatively compared against each other, as well as against the most common video summarization descriptor, i.e., global image histogram, a variant of which was shown in [3] to outperform all competing approaches when no shot boundaries information was available. The two local descriptors (SURF and LMoD) are evaluated with and without the addition of the implemented Dense Trajectories variant. A publicly available annotated dataset [19] and the most common video summarization method, i.e., video frame clustering, are adopted. In all cases, several image channels are being exploited (luminance, hue, edges, optical flow magnitude) in order to simultaneously capture information about the depicted shapes, colors, lighting, textures and motions. A simple, objective evaluation metric is employed for comparing the competing descriptors.

II. A Framework For Describing and Representing Activity Videos

In the proposed approach, each video is assumed to be composed of a temporally ordered sequence $\mathcal{V}$ of N_f video frames, each one being a set $\mathcal{V}_i$ of K matrices $\mathbf{V}_{ik} \in \mathbb{R}^{M \times N}$, where $0 \leq i < N_f$ and $k \in l, h, o, e$. K is the number of available image channels: l stands for luminance, h for color hue, o for optical flow magnitude and e for edge map. Each $\mathbf{V}_{ik}$ is a digitized 8-bit image with a resolution of $M \times N$ pixels.

The presented framework operates by initially computing a set of low-level, L-dimensional description vectors at each $\mathbf{V}_{ik}$. For a given i, a single set of descriptors $\mathcal{D}_i$ is subsequently derived from all image channels, by simply concatenating corresponding vectors computed for different values of k. The correspondence among channels is established in terms of spatial pixel coordinate matching. Each $\mathcal{D}_i$, composed of P_i LK-dimensional, multichannel description vectors, is then transformed into a single histogram feature vector $\mathbf{d}_i$ by following a Bag-of-Features representation approach [12]. That is, all multichannel description vectors from the entire video are clustered into Kc representative groups, called visual words. The set of all cluster centroids is called a *codebook* and c is the codebook size parameter. Each of the P_i vectors in $\mathcal{D}_i$ is subsequently assigned to the nearest visual word, in terms of Euclidean distance. The number of description vectors assigned to each of the Kc clusters is an entry in a Kc-dimensional vector. This vector is followingly transformed into a histogram by L_1-normalization, in order to produce the final Kc-dimensional video frame feature vector $\mathbf{d}_i$. The histogram construction process is repeated for all N_f values of i.

Any type of low-level descriptor can be employed during the first step of the algorithm. This is trivial in the case of local descriptors, such as SIFT or SURF, but not when global descriptors are to be used. In the context of the proposed framework, a variant of the spatial pyramid approach [20] presented in [8] has been adopted for any global descriptor. That is, each $M \times N$ video frame $\mathbf{V}_{ik}$, is partitioned in small blocks of $m \times n$ pixels, where $m < M$ and $n < N$. A description vector is then computed separately for each such block. The process is successively repeated d times, for larger values of m and n, until $m = M$ and $n = N$ during the last iteration. From an implementation point-of-view, this is executed recursively, in a top-down manner, with the image region that is currently being described at each time, subsequently being partitioned into 4 quadrants. These quadrants serve as input blocks to the 4 recursive function calls of the next step. Thus, the total number S of produced description vectors is given by the sum of the first d terms of a geometric progression:

$$S(d) = 1 \cdot 4^0 + 1 \cdot 4^1 + \cdots + 1 \cdot 4^{d-1} = (4^d - 1)/3 \quad (1)$$

In the end, a set of description vectors over various image regions and for various scales is produced, including one truly global description vector (computed over the entire $\mathbf{V}_{ik}$). Available local information is more spatially focused for higher values of d, at the cost of higher computational requirements. In general, however, the main advantage of global image descriptors, i.e., rapid computation [21], is mostly retained with this simple spatial partitioning scheme, in comparison to more complicated alternative approaches, such as image segmentation.

Below, the descriptors used for the evaluation of the proposed framework are presented.

A. Global Descriptors

Global histograms computed in various image channels are the most commonly used feature descriptors for video summarization. For instance, in [6], 16-bin hue histograms derived from the video frame representation in the HSV color space are employed. In the presented framework, a histogram resolution of 16 bins is also adopted in the context of the multichannel, video frame partitioning scheme previously described.

FMoD [8] was also adopted and adapted to our multichannel, video frame partitioning scheme. FMoD operates at each $m \times n$ block by computing one *profile vector* for the horizontal dimension and one for the vertical dimension, through averaging pixel values across block columns / rows, respectively. The result is an n-dimensional and an m-dimensional profile vector. Each of the two vectors is

summarized by their first 4 statistical moments (mean, standard deviation, skewness, kurtosis). The resulting 8-dimensional vector $\mathbf{f}_i = [m_{1H}, m_{2H}, m_{3H}, m_{4H}, m_{1V}, m_{2V}, m_{3V}, m_{4V}]^T$ compactly captures the statistical properties of the block.

In this work, FMoD was extended by adding first-order statistical texture analysis components to the summary of each profile vector, i.e., energy and entropy. Moreover, on top of the horizontal and the vertical profile vector, a third *block vector* is constructed by vectorizing the actual block in row-major order. The same statistical synopsis is also applied on this vector, resulting in an 18-dimensional block description vector $\mathbf{f}_i = [m_{1H}, \cdots, m_{6H}, m_{1V}, \cdots, m_{6V}, m_{1B}, \cdots, m_{6B}]^T$. Using this notation, H, V and B refer to the extracted statistical properties of horizontal profile vectors, vertical profile vectors and block vectors, respectively.

B. Local Descriptors

The most commonly employed local image descriptor is SIFT, with the less computationally costly SURF being a close second choice. Both produce histograms of edge orientations for carefully selected image interest points, in a scale- and rotation-invariant manner. This stems from their design with object recognition tasks in mind, but is not necessarily an ideal approach for the domain of activity video summarization. Since the video description process is not meant to enable successful video classification, but salient key-frame extraction, and given that the subsequent BoF representation step provides (to a degree) several invariances, local descriptors can be of a *holistic* nature, i.e., they would only need to compactly capture major characteristics of untransformed image patches, covering most of (or even the entire) video frame.

In the presented framework, the SURF detector and descriptor [11] was adopted, for reasons of computational speed. Interest point detection occurs on the luminance video frame channel and the detected key-point coordinates are used for SURF description vectors computation on all employed channels.

Additionally, the global, extended FMoD descriptor discussed in Subsection II-A, provided the basis for a novel local descriptor, called Local Moments Descriptor (LMoD). LMoD operates in the manner presented below.

To describe a given image block $\mathbf{B}$ with a dimension of $b \times b$ pixels, $\mathbf{B}$ is recursively partitioned into quadrant sub-blocks using the video frame partitioning scheme of the global descriptor case. For each sub-block, the 18-dimensional description vector of the extended FMoD descriptor is computed and all such vectors are concatenated into an $18S(d)$-dimensional block description vector, where $S(d)$ is given by Equation (1).

Instead of sparsely detecting interest points, as in the case of SIFT or SURF, the luminance channel of the i-th video frame ($\mathbf{V}_{il}$) is densely sampled on a rectangular grid to extract the block centers where LMoD vectors are to be computed, using a sampling step of s pixels. Subsequently, each candidate block is checked for luminance homogeneity, in order to dismiss blocks conveying minimal information. To achieve rapid computation, this is simply implemented using a threshold t_l on the standard deviation of the block luminance.

Dense sampling of interest points allows the background of a depicted activity to be taken into account and complements the holistic nature of LMoD descriptors. As in the case of SURF, description vectors are constructed on all employed channels at the spatial coordinates computed in $\mathbf{V}_{il}$.

C. Activity Descriptor

The multichannel global and local descriptors previously presented are able to describe each video frame in several ways. However, motion information is provided only from the optical flow magnitude image channel and, therefore, is too temporally localized. Since the focus is on activity videos, an additional descriptor may be employed along with local descriptors, which attempts to capture consistent motion in wide temporal windows. This descriptor, called Trajectories, has been adopted and adapted from one component of the state-of-the-art Dense Trajectories description algorithm (designed for activity recognition) [16].

Below, a set of $\mathcal{D}_i$ computed description vectors (corresponding to D_i detected interest points) is assumed for each $\mathbf{V}_{il}$. Trajectories operate by tracking each local interest point along consecutive video frames, using the estimated optical flow magnitude image channel, for a temporal window of T_w frames, in a sliding window approach. Thus, for each local descriptor, a temporally ordered sequence of T_w coordinate pairs (x, y) (referring to horizontal and vertical pixel positions, respectively) is produced, which is equivalent to a set of spatiotemporal coordinate triplets of the form (x, y, t). If a coordinate triplet is shared among different sequences, one sequence is retained and the rest are discarded. A $2T_w$-dimensional vector of spatial displacements is then computed from each sequence, by subtracting the corresponding spatial coordinates among all pairs of subsequent video frames. Static vectors of displacements (derived from interest points with no motion) are eliminated, through comparing their L_1 norm with a threshold t_s. Finally, each retained vector is normalized by the sum of its displacement magnitudes. The result is a *trajectory* description vector $\mathbf{t}_j$, $0 \leq j < P$, where P is the total number of estimated trajectories across the entire video. Each $\mathbf{t}_j$ encodes relative motion direction patterns across a wide temporal window, in a partially scale-invariant manner. For each trajectory starting at video frame b, b is also recorded and, thus, it is trivial to determine which frames $\mathbf{t}_j$ passes through (i.e., which ones are *contained* in it). The starting frame of $\mathbf{t}_j$ is hereafter denoted by b_j.

The set of all trajectory vectors from all video frames is employed to construct a codebook of size c, which is subsequently used to compute a trajectory histogram $\mathbf{h}_i$ per video frame $\mathbf{V}_{il}$. Unlike in the traditional BoF approach, computing $\mathbf{h}_i$ includes a simple weighting scheme: the contribution of each trajectory $\mathbf{t}_j$ is weighted based on the relation between temporal positions b_j and i. That is, the corresponding weight w_j is derived from a discrete Gaussian over the temporal axis with its peak at position i, where each $\mathbf{t}_j$ is assigned to position $b_j + \lceil (T_w/2) \rceil$. Obviously, trajectories not containing position i are completely disregarded. In the end, a c-dimensional trajectory histogram $\mathbf{h}_i$ has been produced for each video frame $\mathbf{V}_{il}$, encoding spatiotemporal activity information.

By employing the approach described above, activity motion descriptions are computed as video features complemen-

tary to local description vectors, in a manner that allows per-frame activity representation.

III. Quantitative Evaluation

A. Evaluation Dataset

In order to experimentally evaluate the proposed framework and descriptors, a subset of the publicly available, annotated IMPART video dataset [19] was employed. It depicts three subjects/actors in two different settings: one outdoor and one indoor. A living room-like setting was set-up for the latter, while two scripts were executed during shooting, prescribing human activities by a single human subject: one for the outdoor and one for the indoor setting. In each shooting session, the camera was static and the script was executed three times in succession, one time per subject/actor. This was repeated three times per script, for a total of 3 indoor and 3 outdoor shooting sessions. Thus each script was executed three times per actor. Three main actions were performed, namely "Walk", "Hand-wave" and "Run", while additional distractor actions were also included and jointly categorized as "Other" (e.g., "Jump-up-down", "Jump-forward", "Bend-forward"). During shooting, the actors were moving along predefined trajectories defined by three waypoints (A, B and C). Summing up, the dataset consists of 6 MPEG-4 compressed video files with a resolution of 720 x 540 pixels, where each one depicts three actors performing a series of actions one after another. The mean duration of the videos is about 182 seconds, or 4542 frames.

The fact that ground truth annotation data provided along with the IMPART dataset describe not key-frames pre-selected by users, as in [6] (which would be highly subjective), but obvious activity segment video frame boundaries, was exploited to evaluate the proposed framework as objectively as possible. Given the results of each summarization algorithm for each video, the number of extracted key-frames derived from actually different activity segments (hereafter called *independent key-frames*) can be used as an indication of summarization success. Therefore, the ratio of extracted independent key-frames by the total number of requested key-frames K, hereafter called *Independence Ratio* (IR) score, is a practical evaluation metric.

B. Experimental Results

The proposed framework and video frame descriptors, as well as a multichannel variant of the global image histogram (without the BoF representation stage) which is popular in the relevant literature (e.g., in [5], [6]), were evaluated on the presented IMPART dataset, using the IR metric and the K-Means++ algorithm [22] for frame clustering, as the main summarization method. Other clustering algorithms have been tested and shown to provide similar results. The method in [23] was employed for optical flow estimation. The Laplace operator was used for deriving the edge map image channel, after 3×3 median filtering for noise suppression.

The number of clusters K, i.e., the number of requested extracted key-frames per video, is a user-provided parameter which controls the grain of summarization. Typically, in clustering-based summarization approaches, K is set proportionally to video length and in accordance with the desired summary consciseness. In order to most effectively compare the different description and representation schemes in terms of the achieved IR score, the actual number Q of different activity segments (known from the ground truth) was used as K for each video. Codebook size c was set to 80, while frame partitioning depth d was set to 6 for global descriptors and to 2 for LMoD. Block dimension, interest point sampling step and luminance dispersion threshold were set to 25 pixels, 20 pixels and 20 standard deviation units, respectively, for LMoD. Interest point tracking temporal window width T_w was set to 15 video frames for Trajectories, as in [16]. The experiments were performed on a high-end PC, with a Core i7 @ 3.5 GHz CPU and 32 GB RAM, while the codebase was developed in C++.

TABLE I: A comparison of the mean IR scores for different video description and representation methods, using K-Means++ summarization.

Method	mean IR
Framework Histogram	0.685
Extended FMoD	0.723
LMoD	0.740
SURF	0.484
LMoD+Trajectories	**0.802**
SURF+Trajectories	0.553
Global Histogram	0.571

TABLE II: A comparison of the mean execution time requirements per-frame for different video description and representation methods.

Method	Mean time (msecs)
Framework Histogram	1077
Extended FMoD	1508
LMoD	1405
SURF	1208
LMoD+Trajectories	2204
SURF+Trajectories	1998
Global Histogram	**706**

Table I presents the IR scores, averaged over the entire employed dataset, that were achieved by the competing approaches. Two global descriptors (Framework Histogram, extended FMoD), two local descriptors (SURF, LMoD) and two composite schemes consisting of the local descriptors and the presented per-frame activity descriptor (SURF+Trajectories, LMoD+Trajectories) were compared. In the last case, the BoF-derived histograms from the local and the activity descriptors were concatenated before frame clustering. Additionally, a traditional 16-bin global image histogram descriptor (omitting the frame partitioning and the BoF representation stages) was employed, for a total of 7 competing approaches. In all cases, all discussed video frame channels (luminance, color hue, optical flow magnitude map, edge map) were exploited through description vector concatenation.

Table II presents the mean required execution times per-frame (in milliseconds), over the entire employed dataset, that were achieved by the competing approaches. These measurements include the time necessary for all description and representation stages for all image channels, as well as the time needed for image channel computation per-frame.

As it can be seen, the proposed framework is outperformed

by the typically used global image histogram only when local SURF descriptors are used, which confirms the findings of [3]. In all other cases, the presented description and representation schemes achieve higher performance, with Extended FMoD being more successful than Framework Histogram and LMoD providing the best results in both metrics. Additionally, the Trajectories per-frame activity descriptor seems to beneficially enrich the informational content of both employed local descriptors (LMoD and SURF), resulting in the combination LMoD+Trajectories being the best choice. Obviously, it is reasonable that activity description aids the summarization of activity videos. Not unexpectedly, this comes at the cost of a threefold increase in required computational time in comparison to the traditional global image histograms. This indicates a typical trade-off between summarization quality and computational requirements, with better performing descriptors being more appropriate for off-line/non real-time applications.

The very low performance of SURF is of high interest, since it validates our assumption that sparsely sampled and highly invariant descriptors designed for recognition tasks are not necessarily suitable for video summarization. This fits with previous results in [3], which indicated that in the absence of clear shot boundary information, global image color histograms produced better results than SIFT and SURF. A local descriptor that is holistic, according to our definition, and densely sampled, i.e., LMoD, outperforms both approaches, possibly because it captures spatial image properties lost in the case of simple global histograms and information content discarded by recognition-oriented local descriptors. Overall, the presented framework seems to be more efficient when employing holistic local descriptors.

IV. Conclusions

We have proposed a consistent video frame description and representation framework based on spatial pyramid partitioning. It accommodates per-frame local, global and activity descriptors, with the goal of assisting successful automated summarization of human activity videos. The framework has been objectively evaluated on a publicly available dataset (IMPART), using the most common video summarization method, i.e., video frame clustering. In all cases, several image channels are being exploited (luminance, hue, edges, optical flow magnitude) in order to simultaneously capture information about the depicted shapes, colors, lighting, textures and motions. In this context, the introduced Extended FMoD, LMoD and Trajectories descriptors (specially adapted novel extensions of pre-existing descriptors) outperform competing approaches, with the LMoD+Trajectories combination proving to be the most effective.

V. Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 316564 (IMPART). This publication reflects only the author's views. The European Union is not liable for any use that may be made of the information contained therein.

References

[1] A. Evans, J. Agenjo, and J. Blat, "Combined 2D and 3D Web-based visualisation of on-set big media data," in *IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 1120–1124.

[2] A. G. Money and H. Agius, "Video summarization: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Representation*, vol. 19, no. 2, pp. 121–143, 2008.

[3] E J.Y. Cahuina and G. C. Chavez, "A new method for static video summarization using local descriptors and video temporal segmentation," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*. 2013, pp. 226–233, IEEE.

[4] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 41, no. 6, pp. 797–819, 2011.

[5] Y. Zhuang, Y. Rui, T.S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *International Conference on Image Processing (ICIP)*. 1998, pp. 866–870, IEEE.

[6] S. E. F. De Avilla, A. P. B. Lopes, A. L. Jr. Luz, and A. A. Araujo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

[7] T. Wan and Z. Qin, "A new technique for summarizing video sequences through histogram evolution," IEEE, 2010, pp. 1–5.

[8] I. Mademlis, N. Nikolaidis, and I. Pitas, "Stereoscopic video description for key-frame extraction in movie summarization," in *European Signal Processing Conference (EUSIPCO)*. 2015, pp. 819–823, IEEE.

[9] J. Li, "Video shot segmentation and key frame extraction based on SIFT feature," in *International Conference on Image Analysis and Signal Processing (IASP)*. IEEE, 2012, pp. 1–8.

[10] D. G. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision (ICCV)*. IEEE, 1999, pp. 1150–1157.

[11] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[12] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *European Conference on Computer Vision (ECCV)*, 2004, pp. 1–2.

[13] Z. Tian, J. Xue, X. Lan, C. Li, and N. Zheng, "Key object-based static video summarization," in *ACM International Conference on Multimedia*, 2011, pp. 1301–1304.

[14] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 82–91, 2006.

[15] W. Fu, J. Wang, L. Gui, H. Lu, and S. Ma, "Online video synopsis of structured motion," *Neurocomputing*, vol. 135, pp. 155–162, 2014.

[16] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by Dense Trajectories," in *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, 2011, pp. 3169–3176.

[17] I. Mademlis, A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Exploiting stereoscopic disparity for augmenting human activity recognition performance," *Multimedia Tools and Applications*, pp. 1–20, 2015.

[18] N. Kourous, A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Video characterization based on activity clustering," in *International Conference on Electrical and Computer Engineering (ICECE)*. IEEE, 2014, pp. 266–269.

[19] H. Kim and A. Hilton, "Influence of colour and feature geometry on multi-modal 3D point clouds data registration," in *International Conference on 3D Vision (3DV)*, 2014, pp. 202–209.

[20] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[21] O.A.B. Penatti, E. Valle, and R. da S. Torres, "Comparative study of global color and texture descriptors for Web image retrieval," *Journal of Visual Communication and Image Representation*, vol. 23, no. 2, pp. 359–380, 2012.

[22] D. Arthur and S. Vassilvitskii, "K-Means++: the advantages of careful seeding," in *Symposium on Discrete Algorithms*. 2007, pp. 1027–1035, Society for Industrial and Applied Mathematics.

[23] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image analysis*, pp. 363–370. Springer, 2003.