Summarization of Human Activity Videos Via Low-Rank Approximation

Ioannis Mademlis†, Anastasios Tefas†, Nikos Nikolaidis† and Ioannis Pitas†⋆

†Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
⋆Department of Electrical and Electronic Engineering, University of Bristol, UK

***Abstract*—Summarization of videos depicting human activities is a timely problem with important applications, e.g., in the domains of surveillance or film/TV production, that steadily becomes more relevant. Research on video summarization has mainly relied on global clustering or local (frame-by-frame) saliency methods to provide automated algorithmic solutions for key-frame extraction. This work presents a method based on selecting as key-frames video frames able to optimally reconstruct the entire video. The novelty lies in modelling the reconstruction algebraically as a Column Subset Selection Problem (CSSP), resulting in extracting key-frames that correspond to elementary visual building blocks. The problem is formulated under an optimization framework and approximately solved via a genetic algorithm. The proposed video summarization method is being evaluated using a publicly available annotated dataset and an objective evaluation metric. According to the quantitative results, it clearly outperforms the typical clustering approach.**

***Keywords*—Video Summarization, Sparse Dictionary Learning, Genetic Algorithm**

I. Introduction

In recent years, the need for succinct presentation of digital videos depicting human activities has increased exponentially. Data from surveillance cameras or professional capture sessions are two such cases where an automated algorithmic solution would greatly benefit the end-users, allowing them rapid browsing, analysis, annotation and archiving of lengthy video footage, while significantly reducing storage requirements. *Video summarization* addresses this problem by generating compact versions of a video stream, after having determined its most informative and representative content [1]. *Static summarization* algorithms typically extract a set of salient video frames, i.e., *key-frames* that represent the entire video content. They are contrasted with *dynamic summarization* methods, where a *video skim* is being constructed as a sequence of short video *key-segments* concatenated in the correct temporal order, thus forming a meaningfully shortened version of the original stream. This work deals with the problem of static video summarization, i.e., key-frame extraction, and skim construction is not addressed.

Typically, information is extracted by analysing the available modalities (visual, audio or textual) to detect high-level semantic content, e.g. depicted objects or events, as well as computing low-level features from the video stream. To accomplish this task, each video frame is first described by low-level image features, with the most commonly employed frame descriptor being variants of global joint image histograms in the HSV color space [2] [3].

In most of the relevant literature, summarization is implicitly defined as a frame sampling problem, with systematic sample acquisition methods being presented that try to simultaneously satisfy several heuristic criteria, such as compactness (lack of content redundancy in the selected key-frames / key-segments), outlier inclusion (selection of atypical key-frames / key-segments) and coverage (representation of the entirety of the original video in the produced summary). The traditional summarization method derived from this heuristic definition is video frame clustering, e.g., with the frames closest to the estimated cluster centroids being selected as key-frames. The number of clusters may depend on the video length [2].

Various similar summarization approaches have also been proposed, implicitly obeying the aforementioned heuristic criteria: e.g., a computational geometry-based approach [4] that results in key-frames equidistant to each other in the sense of video content, or a fast method which selects as key-frames the video frames that locally maximize an aggregate intra-frame difference (computed using color features) [5]. However, clustering still dominates the relevant literature due to its simplicity, suitability to the problem and relatively low computational requirements. In many cases, information about the way a video is naturally segmented into shots (e.g., in movies [6]) is also exploited to assist the summarization process [7] [3] [2] [8], e.g. by applying clustering at shot-level. Typically, the extracted key-frame set is pruned in a refinement post-processing stage. The remaining key-frames are temporally ordered to produce a meaningful summary.

The above described approaches can be applied to generic video, while methods exploiting video type-specific information have also been proposed. In surveillance videos, temporal segmentation (shot boundaries detection [9]) 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 [10]. However, this is not a useful approach in the very similar scenario of raw videos from professional capture sessions (e.g., television or film production), where also the camera is static and natural segmentation into shots is absent, since the preferred summarization goal would be to select one key-frame per depicted activity.

In [11] and [12] the video summarization problem is formulated in terms of sparse dictionary learning, with extracted key-frames enabling optimal reconstruction of the original video from the selected dictionary. Such an approach implies an interesting and formal definition of a video summary, as the set of key-frames that can linearly reconstruct the full-length video in an algebraic sense. However, the conciseness of the summary is only enforced via optimization using a sparsity constraint, with no guarantees that such a process will actually converge to a small number of key-frames. Thus, compactness is not assured.

Our paper, following in this line of work, attempts to overcome

this limitation and provide a novel reconstruction-based algebraic method where the number of key-frames is a fixed, user-provided parameter, as in typical clustering approaches. To this end, the Column Subset Selection Problem (CSSP) is employed for problem modelling and solved using a genetic algorithm. Thus, the proposed method is able to extract a set of key-frames constituting elemental visual building blocks of the original video sequence, implicitly defining the summary as a subset of the video frames from which the entire full-length video may be linearly reconstructed. To our knowledge, the CSSP has not been employed before in the context of a key-frame extraction algorithm.

II. Video Summarization Based on the Column Subset Selection Problem

The first step of video summarization is video frame description. In the proposed approach, each video is assumed to be composed of a temporally ordered sequence of N_f video frames of dimension $M \times N$, each one being a set 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.

A global and a local descriptor were separately employed and compared: global, 16-bin video frame intensity histogram, as well as visual word histograms based on SURF descriptors [13] and a Bag-of-Features representation scheme [14]. Intensity histograms were selected due to their prevalence in video summarization literature, while SURF descriptors due to their great performance at a relatively low computational cost in object recognition applications [13]. In the first case, for the i-th frame, the histograms are being separately computed on each image channel and then concatenated. In the second case, a single set of descriptors $\mathcal{D}_i$ is derived by simply concatenating corresponding 128-dimensional SURF vectors separately computed on the available channels. The vector correspondence between channels is established in terms of spatial pixel coordinate matching, while the interest points are initially detected solely on $\mathbf{V}_{il}$, i.e., on luminance. Each $\mathcal{D}_i$, composed of P_i $128K$-dimensional description vectors, is then transformed into a single Kc-dimensional BoF visual word histogram $\mathbf{d}_i$ [14], where c is a codebooks size parameter. The adoption of the BoF approach was motivated by its proven suitability for the representation of human activities, since it discards most of the spatial information and thus provides partial invariance to changes in camera viewpoint, number of human subjects, scale, rotation and occluded object parts [15].

Human activity videos are mainly composed of elementary visual building blocks assembled in several combinations. Given the above video description strategy, this is expressed with each video being represented as a histogram matrix $\mathbf{D}$, where several columns constitute linear combinations of other columns. Thus, the summarization objective is for the estimated summary $\mathbf{C}$ to mainly contain columns that form a set of linearly independent basis vectors, spanning the space of all columns in $\mathbf{D}$. In this sense, C will tend to be able to reconstruct the original matrix D in a manner well-suited to the task at hand, ideally extracting key-frames representative of all the depicted human activities.

Therefore, the proposed method models key-frame extraction as a matrix Column Subset Selection Problem (CSSP) [16], which, to our knowledge, has not been attempted before. Below, the CSSP is briefly discussed. Assuming a low-rank $Kc \times N_f$ matrix $\mathbf{D}$ and a parameter $C < N_f$, CSSP consists in selecting a subset of exactly C columns of $\mathbf{D}$, which will form a new $Kc \times C$ matrix $\mathbf{C}$ that captures as much of the information contained in the original matrix as possible. The goal is to construct a matrix $\mathbf{C} \in \mathbb{R}^{Kc \times C}$ such that the quantity

$$\|\mathbf{D} - (\mathbf{C}\mathbf{C}^{+})\mathbf{D}\|_F \tag{1}$$

is minimized. In the above, $\|\cdot\|_F$ is the Frobenius matrix norm and $\mathbf{C}^{+}$ is the pseudoinverse of $\mathbf{C}$. Thus, the goal is to minimize the reconstruction error between the entire video $\mathbf{D}$ and the projection of $\mathbf{D}$ onto the span of the C columns contained in the summary $\mathbf{C}$. If $\mathbf{C}$ was a full-rank matrix, then $\mathbf{C}\mathbf{C}^{+}$ would equal the identity matrix and the reconstruction error would be 0. Thus, minimizing Equation (1) is equivalent to finding a subset matrix $\mathbf{C}$ that is as close to full-rank as possible.

CSSP is an obvious choice for mathematically modelling a feature selection process as an optimization problem. It can be optimally solved by exhaustive search in $\mathcal{O}(N^C)$ time [16], which clearly is a very impractical approach. Thus, approximate algorithms with lower computational complexity have been presented in the relevant literature, with the goal of finding a suboptimal but acceptable solution.

In [17], a genetic approach is successfully employed for the approximate solution of the CSSP, by directly using Equation (1) as a fitness function. The method is evaluated on several small, randomly generated matrices and is shown to produce good results for a fixed small value of C. In this work, the same approach was adopted and adapted into the proposed algorithm.

Due to the nature of the CSSP, there is no need for a regularizing function $R(\mathbf{C})$, like the one in [12]. The degree of summary compactness and conciseness is directly regulated by a strict, user-provided parameter C, as in most commonly employed clustering-based summarization methods. The desired solution is a set of matrix column indices with cardinality equal to C. Since $\mathbf{D} \in \mathbb{R}^{Kc \times N_f}$, for the k-th such index with an assigned value g_k the following hold:

$$k \in \mathbb{N}, \quad k \in [1, \cdots, C]. \tag{2}$$

$$g_k \in \mathbb{N}, \quad g_k \in [1, \cdots, N_f]. \tag{3}$$

A genetic algorithm is employed to approximate an optimal solution [17]. Each candidate/chromosome is encoded in the form of a sequence of column indices sorted in increasing order. Every such chromosome is of length C and population size is N. Roulette selection at each iteration is adopted as the mating pool formation strategy. Assuming $fit(l)$ is the evaluated fitness of $\mathbf{h}^l$, i.e., the l-th candidate in the current population, this method assigns a selection probability $p^l_{sel} = fit(l)/\sum_{m=1}^{N} fit(m)$ to the l-th chromosome. Below, the value assigned to the k-th gene of a chromosome $\mathbf{h}^l$ is denoted by $\mathbf{h}^l_k$.

An order-preserving variant of 1-point crossover [17] is utilized as the main genetic operator. Specifically, in order to combine parent chromosomes $\mathbf{h}^l$ and $\mathbf{h}^m$, a random position k is selected as crossover point and is inspected for suitability. k is considered to be suitable as a crossover point, if the following condition holds:

$$(\mathbf{h}^l_k < \mathbf{h}^m_{k+1}) \wedge (\mathbf{h}^m_k < \mathbf{h}^l_{k+1}). \tag{4}$$

This constraint ensures that both offspring will be valid candidates, containing properly ordered matrix column indices. In case Equation (4) does not hold for position k, a different position is selected and

inspected. This process continues until either a suitable crossover point has been detected, or all possible positions have been deemed unsuitable. In the former case, crossover is applied and the two parent chromosomes are replaced by their offspring. In the latter case, each of the implicated chromosomes is passed unaltered to the population of the next generation with probability p_{sel}^{l} or p_{sel}^{m}, respectively. If $\mathbf{h}^{l}$ or $\mathbf{h}^{m}$ is not being retained, it is replaced in the next generation by a copy of the fittest current candidate $\mathbf{h}^{n}$ with probability p_{sel}^{n}. If $\mathbf{h}^{n}$ is also not selected for retention, the process continues with the second fittest of the current candidates, and so on, until a chromosome has been selected.

An order-preserving variant of mutation [17] is employed as the second genetic operator. Specifically, the k-th gene of a chromosome $\mathbf{h}^{n}$, with an assigned valued of $\mathbf{h}_{k}^{n}$, is randomly selected and replaced by a value determined by the neighbouring genes, according to Equation (5):

$$\mathbf{h}_{k}^{n} = \begin{cases} rand(0, \mathbf{h}_{k+1}^{n}), & \text{if} \quad k = 1 \\ rand(\mathbf{h}_{k-1}^{n}, \mathbf{h}_{k+1}^{n}), & \text{if} \quad k \in (1, C) \\ rand(\mathbf{h}_{k-1}^{n}, N_f + 1), & \text{if} \quad k = C. \end{cases} \tag{5}$$

where $rand(a, b)$ uniformly selects a random integer from the interval (a, b). Although this operator ensures a proper ordering of the indices, it has no effect when $\mathbf{h}_{k-1}^{n}$, $\mathbf{h}_{k}^{n}$ and $\mathbf{h}_{k+1}^{n}$ are successive integers.

The matrix column indices encoded in the evaluated chromosome $\mathbf{h}^{n}$ give rise to the matrix $\mathbf{C}_{n}$, composed of a subset of the columns in $\mathbf{D}$. Thus, the fitness function that needs to be maximized is expressed as:

$$fit(\mathbf{h}^{n}) = \|\mathbf{D} - (\mathbf{C}_{n}\mathbf{C}^{+}{}_{n})\mathbf{D}\|_{F}^{-1}. \tag{6}$$

The method may be easily extended to accommodate additional desired summary properties, through proper manipulation of the employed fitness function. Additionally, an interesting research avenue would be a way to evaluate summarization results for different values of the parameter C, i.e., the desired key-frame set cardinality, since ground truth is typically not available. This resembles the problem of selecting a proper K in K-Means clustering. An obvious route to tackle this problem is to run the algorithm for multiple consecutive values of C and construct a signal with the corresponding CSSP reconstruction errors. It is reasonable to expect the error to steadily decrease for larger values of C. Then, a proper value for C may be identified at the point where the signal's derivative drops below a threshold. Less obvious and more efficient approaches to this problem could be explored in future research.

III. Evaluation

In order to experimentally evaluate the proposed method, a subset of the publicly available, annotated IMPART video dataset [18] 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 action 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. Sample video frames of the dataset are shown in Figure 1.

Fig. 1. *Example frames from the IMPART video dataset. The respective activities are "Run" (top left), "Walk" (bottom left), "Jump" (top right) and "Hand-wave" (bottom right).*

TABLE I. A comparison of the mean IR scores for different video description/representation and summarization methods.

Method	K-Means++	CSSP
Global Histogram	0.571	**0.636**
SURF	0.484	**0.534**

Ground truth annotation data provided along with the IMPART dataset do not describe key-frames pre-selected by users, as in [2] (which would be highly subjective), but obvious activity segment frame boundaries. This fact 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.

The proposed method and the K-Means++ algorithm [19] for frame clustering were objectively evaluated and contrasted using the IMPART dataset and the IR metric. The fast OpenCV [20] implementations of the method in [21] and of the SURF detector and descriptor were employed for optical flow estimation and local video frame description, respectively. In all video frames, the Laplace operator was used for deriving the edge map image channel, after median-filtering for noise suppression.

TABLE II. A comparison of the mean execution time requirements per-frame (in milliseconds) for different video description/representation and summarization methods.

Method	K-Means++	CSSP
Global Histogram	**706**	1119
SURF	**1208**	1789

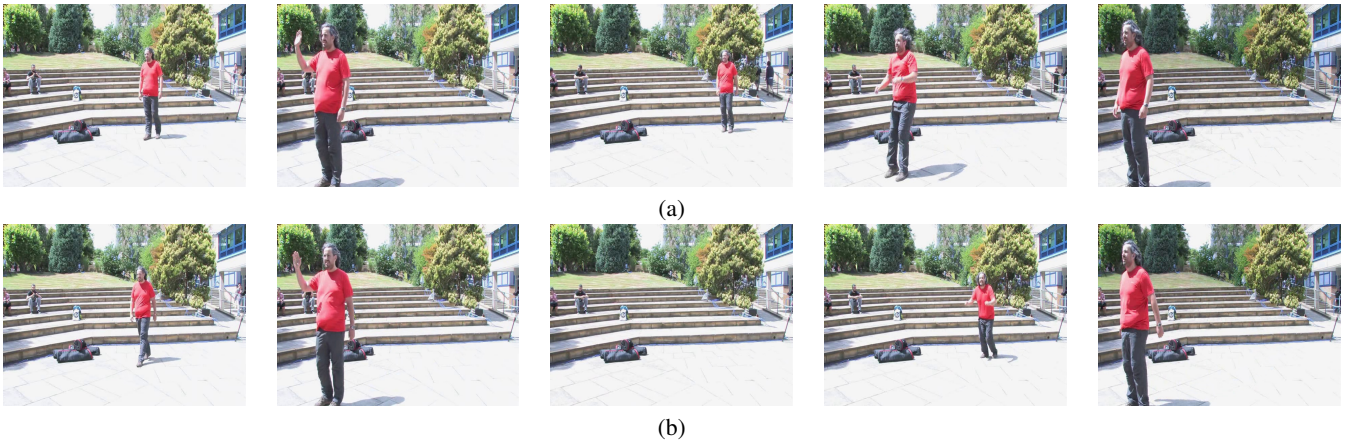

Fig. 2. Key-frames extracted on a sample of the IMPART dataset, using a) K-Means++ and b) the proposed CSSP-based method.

A crucial, user-provided parameter controlling the grain of summarization is the desired number of clusters K and of the columns C in the summary matrix $\mathbf{C}$, in clustering and in the proposed summarization method, respectively. It corresponds to the number of requested extracted key-frames per video. The actual number Q of different activity segments (known from the ground truth) was used both as K and C for each video. Codebook size c was set to 80, while the following parameters were used for the genetic algorithm: the maximum number of generations was set to 100, the population size L was set to 200, the crossover rate was set to 0.9, the mutation rate was set to 0.005 and the elitism rate was set to 10%. 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 presents the IR scores, averaged over the entire employed dataset, that were achieved by the two competing approaches, using the. two employed video description schemes (global intensity histograms and SURF-based visual word histograms). 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, representation and summarization stages for all image channels, as well as the time needed for image channel computation per-frame.

Figures 2a,b depict the key-frames extracted using the K-Means++ algorithms and the proposed CSSP-based method, respectively, on a short sample of the IMPART dataset, composed of 5 activity segments: two "Walk", one "Run", one "Hand-wave" and one "Other" action (a "Jump Forward"). K and C were set to 5, while the global image histograms description scheme was employed. By visual inspection, clustering seems to produce a key-frame set with greater redundancy (two frames are almost identical), while the proposed method apparently decomposes the video into elemental visual word subsets, including a blank key-frame depicting only the static background. It is interesting that the "Run" segment is not captured by either method, which may be attributed to its high similarity to the "Walk" segments. It is reasonable that a more elaborate video description scheme is necessary to overcome this limitation.

As it can be seen, local SURF descriptors are outperformed by the more common and faster global image histograms, which confirms the findings of [22] that in the absence of clear shot boundary information, global image color histograms produce better results than SIFT and SURF. This suggests that sparsely sampled and highly invariant descriptors designed for recognition tasks are not necessarily suitable for video summarization. Regarding the competing summarization approaches, it is evident that the proposed method is quantitatively better than the established clustering technique, in terms of the IR metric. However, this comes at the cost of higher computational requirements: it demands approximately 1.5 times the runtime of the clustering approach.

IV. Conclusions

A matrix reconstruction-based method for summarization of videos depicting human activities was presented, that guarantees desired conciseness through fixed key-frame set cardinality. The novelty lies in modelling the problem as a Column Subset Selection Problem (CSSP), with the extracted key-frames corresponding to elementary visual building blocks that may linearly reconstruct the original video. This was formulated under an optimization framework and approximately solved via a genetic algorithm. The proposed video summarization method was evaluated using a publicly available annotated dataset and an objective evaluation metric. According to the quantitative results, it clearly outperforms the typical clustering approach, while previous findings regarding the suitability of simple global image histograms, in contrast to recognition-oriented local image descriptors, to the task of video summarization were validated.

Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement numbers 316564 (IMPART) and 287674 (3DTVS).

References

[1] 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.

[2] 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.

[3] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, "STIMO: STIll and MOving video storyboard for the web scenario.," *Multimedia Tools and Applications*, vol. 46, no. 1, pp. 47–69, 2010.

[4] C. Panagiotakis, A. Doulamis, and G. Tziritas, "Equivalent key frames selection based on iso-content distance and iso-distortion principles," in *International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2007, pp. 29–29.

[5] N. Ejaz, T. B. Tariq, and S. W. Baik, "Adaptive key frame extraction for video summarization using an aggregation mechanism," *Journal of Visual Communication and Image Representation*, vol. 23, no. 7, pp. 1031–1040, 2012.

[6] 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.

[7] 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.

[8] G. Guan, Z. Wang, K. Yu, S. Mei, M. He, and D. Feng, "Video summarization with global and local features," in *International Conference on Multimedia and Expo Workshops (ICMEW)*, 2012, pp. 570–575.

[9] 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.

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

[11] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 66–75, 2012.

[12] S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. D. Feng, "Video summarization via minimum sparse reconstruction," *Pattern Recognition*, vol. 48, no. 2, pp. 522–533, 2015.

[13] 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.

[14] 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.

[15] I. Mademlis, A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Exploiting stereoscopic disparity for augmenting human activity recognition performance," *Multimedia Tools and Applications*, vol. 75, no. 19, pp. 11641–11660, 2016.

[16] C. Boutsidis, M. W. Mahoney, and P. Drineas, "An improved approximation algorithm for the Column Subset Selection Problem," in *Symposium on Discrete Algorithms*. 2009, pp. 968–977, Society for Industrial and Applied Mathematics.

[17] P. Kromer, J. Platos, and V. Snasel, "Genetic algorithm for the column subset selection problem," in *IEEE Complex, Intelligent and Software Intensive Systems (CISIS)*, 2014, pp. 16–22.

[18] 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.

[19] 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.

[20] G. Bradski, A. Kaehler, and V. Pisarevsky, "Learning-based computer vision with Intel's open source computer vision library," *Intel Technology Journal*, vol. 9, no. 2, pp. 119–130, 2005.

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

[22] 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.