Summarization of Human Activity Videos Using a Salient Dictionary

Ioannis Mademlis†, Anastasios Tefas† and Ioannis Pitas†⋆
†Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
⋆Department of Electrical and Electronic Engineering, University of Bristol, UK

***Abstract*—Video summarization has become more prominent during the last decade, due to the massive amount of available digital video content. A video summarization algorithm is typically fed an input video and expected to extract a set of important key-frames which represent the entire content, convey semantic meaning and are significantly more concise than the original input. The most wide-spread approach relies on video frame clustering and extraction of the frames closest to the cluster centroids as key-frames. Such a process, although efficient, offloads the burden of semantic scene content modelling exclusively to the employed video frame description/representation scheme, while summarization itself is approached simply as a distance-based data partitioning problem. This work focuses on videos depicting human activities (e.g., from surveillance feeds) which display an attractive property, i.e., each video frame can be seen as a linear combination of elementary visual words (i.e., basic activity components). This is exploited so as to identify the video frames containing only the elementary visual building blocks, which ideally form a set of independent basis vectors that can linearly reconstruct the entire video. In this manner, the semantic content of the scene is considered by the video summarization process itself. The above process is modulated by a traditional distance-based video frame saliency estimation, biasing towards more spread content coverage and outlier inclusion, under a joint optimization framework derived from the Column Subset Selection Problem (CSSP). The proposed algorithm results in a final key-frame set which acts as as salient dictionary for the input video. Empirical evaluation conducted on a publicly available dataset suggest that the presented method outperforms both a baseline clustering-based approach and a state-of-the-art sparse dictionary learning-based algorithm.**

Keywords*—*Video Summarization, Column Subset Selection Problem, Dictionary Learning, Video Saliency

I. Introduction

The volume of video data available in digital form has grown exponentially during the last two decades. As a result, the need has arisen for concise video representation in a variety of tasks. For instance, on-line browsing in video repositories, video retrieval or surveillance feed synopsis can be facilitated by having available subsets of the original video content, carefully selected so as to balance content coverage, semantic representativeness and conciseness. Thus, automated *video summarization* is defined as the problem of properly selecting and representing the most suitable subset of an input video in a way that respects these constraints. Two types of video summaries dominate the literature: sets of still video frames that have been extracted as independent *key-frames*, i.e., *static summaries*, and short video *key-segments* concatenated in temporal order, i.e., *dynamic summaries* or *video skims*. Typically, a key-segment is centered on a video frame pre-identified as key-frame, therefore static key-frame extraction is the decisive component of a video summarization pipeline.

Video frames are initially described by low-level global or local image descriptors. In general, the most commonly employed frame descriptors are variants of global joint image histograms in the HSV color space [1] [2] [3]. Moreover, dimensionality reduction on such color histogram vectors has been attempted [4], in order to decrease the computational cost of the subsequent video summarization. In [5], the low-level Frame Moments Descriptor (FMoD) is introduced that is designed to compactly capture statistical characteristics of several low-level image channels, both in global and in various local scales. In [6], a local variant of FMoD is shown to outperform competing descriptors for a specific video summarization task. In a few cases [7] [8] [9] [10], local image region descriptors, such as Scale-Invariant Feature Transform (SIFT) [11], or Speeded-Up Robust Features (SURF) [12] have been employed for video description, using the popular Bag-of-Features (BoF) representation model [13]. Such advanced image descriptors, as shown by their success in object recognition and image retrieval tasks, are able to convey (to a degree) mid-level semantic scene information (e.g., parts of depicted scene objects). Although this is important for proper summarization, it is achieved by infusing them per constructionem with a high degree of invariance to various image transformations. Such a property is not necesserilly beneficial for a video summarization task, since it contradicts accurate video frame description [6].

In order to extract key-frames, the video frame descriptors are typically processed by clustering algorithms to create video frame groups, under the assumption that video shooting focuses more on important video frames [2]. The number of clusters may depend proportionally on the video length [1]. Subsequently, a set of video frames that are closest to each video frame cluster centroid are selected as key-frames. In many cases, shot detection [14] (e.g., in movies [5]) is also exploited to assist the summarization process [1] [3] [15] [16], e.g., by applying clustering at shot-level. Typically, a percentage of the extracted key-frames is filtered out and the remaining ones are presented in the same temporal order as the one of the original video to produce a summary. Despite the prevalence of clustering-based methods, various other summarization approaches have also been proposed over the

years, implicitly attempting to satisfy criteria such as outlier inclusion, good content coverage and compactness (defined here as lack of redundancy over the key-frame set). However, a specific drawback can be repeatedly identified in most of these techniques: scene semantics are ignored by the summarization algorithm itself, which usually models summarization simply as a distance-based data partitioning problem. The burden of semantics extraction is offloaded solely to the employed video description/representation method.

For instance, a computational geometry-based approach has been proposed in [17], that results in key-frames that are equidistant to each other in the image feature domain. A fast method producing key-frames that locally maximize an aggregate intra-frame difference, which is computed using low-level color features, was employed in [18]. In [19] [20], key-frame extraction is formulated as the selection of video frames that maximize the coverage of various local image features. By computing such a coverage, a global representation score is estimated on a per-frame basis. The local maxima of the resulting video representation curve lead to key-frame extraction. The resulting static summaries consist of key-frames that contain as many representative visual elements as possible. Therefore, outlying video frames are entirely disregarded and content coverage, in the sense of intra-frame packing of multiple local visual details recurring in the video, is promoted as an absolute criterion of proper video summarization.

In contrast, in [21], [22] 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 being the set of key-frames that can linearly reconstruct the full-length video in an algebraic sense. However, the video frame descriptions rely on global image features, de-emphasizing local image details [21], [22]. Furthermore, the outliers are entirely disregarded. Additionally, 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, succinctness and outlier inclusion are not assured.

In this paper, a static summarization algorithm for videos depicting human activities (e.g., from a surveillance feed) is proposed. Such videos are typically filmed with a static camera and no shot cut boundaries are clearly discernible. However, they display an attractive property, i.e., each video frame can be seen as a linear combination of elementary visual words (i.e., basic activity components). Thus, the presented algorithm attempts to extract from an input video the video frames mostly resembling a *salient dictionary*. The dictionary component, meant to guarantee summary conciseness and compactness, inherently operates within the constraints imposed by video semantics, through identification of the video frames containing only the elementary visual building blocks. Thus, semantics extraction is integrated into the video summarization process itself. Dictionary construction is modulated by the saliency component, meant to ensure outlier inclusion and broad content coverage, which operates in the traditional inter-frame distance-based manner. The entire process is formulated within a joint optimization framework derived from the Column Subset Selection Problem (CSSP). Empirical evaluation conducted in a publicly available dataset suggests that our algorithm outperforms both a baseline clustering approach [1] and a state-of-the-art sparse dictionary learning method [22].

II. Video Summary as a Salient Dictionary

A. Problem Definition and Modelling

An input video composed of N_f frames is represented as a matrix $\mathbf{D} \in \mathbb{R}^{V \times N_f}$. Each column vector $\mathbf{d}_j, 0 \leq i < N_f$, describes a video frame. Moreover, we assume that the desired summary is a matrix $\mathbf{C} \in \mathbb{R}^{V \times C}$, $C \leq N_f$ containing an ordered set of video key-frames. Its columns are indicated indicated by a binary-valued frame selection vector $\mathbf{s} \in \mathbb{N}^{N_f}$.

Human activity videos are mainly composed of elementary visual building blocks assembled in several combinations, thus $\mathbf{D}$ is assumed to be low-rank. The proposed algorithm can be seen as the simultaneous optimization of two components: the first one (the "dictionary") represents the ability of the extracted key-frame set to linearly reconstruct the original full-length video, while the second one (the "saliency") represents the saliency of the extracted key-frame set, i.e., the degree to which its elements are distinct with regard to the complete video frame set, thus more likely to attract viewer attention.

The Column Subset Selection Problem (CSSP) [23] has been selected for modelling the dictionary component, due to its ability to directly estimate a full-rank dictionary composed of unaltered columns of the original matrix, a property well-suited to the problem of key-frame extraction. Given $\mathbf{D}$ and a parameter $C < N_f$, CSSP consists in selecting a subset of exactly C columns of $\mathbf{D}$, which will form a new $V \times C$ matrix $\mathbf{C}$ that captures as much of the information contained in the original matrix as possible. The goal is to construct a matrix $\mathbf{C} \in \mathbb{R}^{V \times C}$ such that the quantity:

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

is minimized. $\|\cdot\|_F$ is the Frobenius matrix norm and $\mathbf{C}^{+}$ is the pseudoinverse of $\mathbf{C}$. Thus, the approximation of $\mathbf{D}$ by the smaller matrix $\mathbf{C}$ is expressed in a projection sense: $\mathbf{C}\mathbf{C}^{+}$ acts as a projection matrix onto the span of the C columns contained in $\mathbf{C}$. Minimizing norm (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. As exhaustive search requires $\mathcal{O}(N^C)$ time [23], approximate algorithms with lower computational complexity have been presented, with the goal of finding a suboptimal but acceptable solution. A numerical algorithm, based on the Singular Value Decomposition (SVD), was adopted from [23] and modified for the purposes of the this work.

Due to the nature of the CSSP, there is no need for a regularizer enforcing sparsity on $\mathbf{s}$, like the one in [22]. The degree of summary conciseness is directly regulated by a strict, user-provided parameter C.

Intuitively, the dictionary component alone will tend to favour video frames solely containing common, elementary

visual building blocks of the entire video, which facilitate the reconstruction process. These include not only video frames that are representative of the depicted human actions, but also uninteresting video frames which do not contribute to discrimination among actions (e.g., emphasizing recurring static background or human body poses common to multiple actions). Additionally, outlier video frames which do not contribute to the reconstruction, may be excluded.

Thus, the saliency component must be considered, in order to ensure outlier inclusion. In the presented algorithm, a simple approach was selected for fast outlier detection, adapted from the spatial, intra-frame component of the saliency estimation algorithm presented in [24]. For the purposes of this work, saliency values are assigned to entire video frames, instead of video frame blocks, and spatial distance between the latter ones is replaced by temporal distance between video frames.

We define the fully connected, undirected, weighted distance graph $\mathcal{D} = \{\mathcal{N}, \mathcal{E}\}$ derived from the matrix $\mathbf{D}$, where $\mathbf{n}_i \in \mathcal{N}, 0 \leq i < N_f$, $\mathbf{e}_j \in \mathcal{E}, 0 \leq j < N_f(N_f-1)/2$. Each vertex $\mathbf{n}_i$ corresponds to a column $\mathbf{d}_i$ in $\mathbf{D}$, i.e., a video frame, and each edge $\mathbf{e}_j$ is weighted by the Euclidean distance between its two incident vertices/video frames, namely, the distance between the corresponding columns in $\mathbf{D}$, normalized by the temporal distance between the two vertices. The degree $deg(\mathbf{n}_i)$ of each vertex, i.e., the sum of the weights of all its incident edges, is employed as a measure of video frame saliency.

Thus, the pre-computed saliency of a column $\mathbf{d}_i$, i.e., the i-th entry of per-frame saliency vector $\mathbf{p}$, is given by:

$$\mathbf{p}_i = deg(\mathbf{n}_i) = \sum_{j=0}^{N_f-1} \left(\frac{\|\mathbf{d}_i - \mathbf{d}_j\|_2}{1+|i-j|} \right). \quad (2)$$

The desired solution is the binary-valued video frame selection vector $\mathbf{s} \in \{0,1\}^{N_f}$ and the actual video summary $\mathbf{C} \in \mathbb{R}^{V \times C}$ constructed based on $\mathbf{s}$ ($\|\mathbf{s}\|_1 = C$).

Given the above, the proposed algorithm implicitly optimizes the following criterion:

$$\min_{\mathbf{s}} : \|\mathbf{D} - \mathbf{C}\mathbf{C}^{+}\mathbf{D}\|_F - \alpha c \mathbf{s}^T \mathbf{p}, \quad (3)$$

where $\alpha \in [0,1]$ is a user-provided parameter regulating the contribution of the saliency component and c is a scaling factor to bring per-video frame saliency value down to the scale of the dictionary component. The entries of $\mathbf{p}$ are a priori given by (2).

B. Solving the Optimization Problem

The proposed algorithm extends a state-of-the-art SVD-based method for solving the CSSP [23]. The method operates in two stages. First, approximately $C\log C$ columns are randomly sampled from matrix $\mathbf{D}$. Sampling follows a probability distribution p_i, constructed based on information coming from the top-C right singular subspace of $\mathbf{D}$, which is spanned by the columns of the SVD-provided matrix $\mathbf{V}_C \in \mathbb{R}^{N_f \times C}$:

$$p_i = \|(\mathbf{V}_C)_i\|_2^2 / C. \quad (4)$$

p_i is the probability of selecting the i-th column of $\mathbf{D}$ and $(\mathbf{V}_C)_i$ denotes the i-th row of $\mathbf{V}_C$.

In the second stage, exactly C columns are deterministically selected from the sample. The method provides a good upper bound for the reconstruction error deviation from that of the best C-rank approximation of $\mathbf{D}$ (denoted by $\mathbf{D}_C$):

$$\|\mathbf{D} - \mathbf{C}\mathbf{C}^{+}\mathbf{D}\|_F \leq \mathcal{O}\left(Clog^{1/2}C\right) \|\mathbf{D} - \mathbf{D}_C\|_F, \quad (5)$$

with probability at least 0.7. The algorithm runs in $\mathcal{O}(min\{VN_f^2, V^2N_f\})$ [23].

In order to adapt the original CSSP method to the proposed algorithm, matrix $\mathbf{D}$ is modified in the following manner:

$$\hat{\mathbf{D}} = (1-\alpha)\mathbf{D} + \alpha \mathbf{D}\left(\mathbf{diag}(\mathbf{n})\mathbf{diag}(\mathbf{p})\right), \quad (6)$$

where $\mathbf{n} \in \mathbb{R}^{N_f}$ is a vector containing normalization coefficients, so as to map the pre-computed saliency factors to the interval $[0,1]$. In $\hat{\mathbf{D}}$, less salient columns (corresponding to less salient video frames) have been scaled down to a degree directly proportional to their saliency and to the provided saliency contribution parameter α. Subsequently, the algorithm in [23] is applied on $\hat{\mathbf{D}}$, in order to obtain the desired summary.

C. Video Description and Representation

Matrix $\mathbf{D}$ contains the original video data representation. In the context of this work, the typical Bag-of-Features (BoF) aggregate representation was applied and, thus, each column vector of $\mathbf{D}$ is a BoF histogram corresponding to a video frame. The description vectors that were aggregated into the BoF histograms were derived from the LMoD+Trajectories description process, which was shown in [6] to outperform competing descriptors in the task of summarizing human activity videos. The selected descriptors capture both low-level spatial scene properties across several image channels (luminance, color hue, optical flow magnitude, edge map) and mid-level spatiotemporal semantic human activity properties. Their per-video frame aggregation under a BoF scheme inherently represents each video frame as a histogram of elementary visual words, combined in different ways at each video frame, thus facilitating the dictionary learning component of the proposed algorithm.

III. Quantitative Evaluation

In order to empirically evaluate the proposed algorithm, a subset of the publicly available, annotated IMPART video dataset [25] 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 one, while two scripts were executed during shooting, prescribing human actions 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

TABLE I. A COMPARISON OF THE MEAN IR SCORES FOR DIFFERENT SUMMARIZATION METHODS.

Proposed	[1]	[22]
0.872	0.571	0.802

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 $\times$ 540 pixels, each depicting three actors performing a series of actions one after another. The mean duration of the videos is about 182 seconds, or 4542 frames.

Ground truth annotation data were provided along with the IMPART dataset, describing obvious action video segment frame boundaries, instead of key-frames pre-selected by users, as in [1] (which would be highly subjective). This was exploited to evaluate the proposed summarization framework as objectively as possible, in a manner similar to [6]. Given the results of each summarization algorithm for each video, the number of extracted key-frames derived from actually different action 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.

Using the above described dataset and peformance metric, the proposed algorithm was compared against a baseline clustering approach to video summarization [1] and a state-of-the-art sparse dictionary learning method [22]. α (saliency contribution coefficient) was set to 0.25. The fast OpenCV [26] implementation of the method in [27] was employed for optical flow estimation. In all video frames, the Laplace operator was used for deriving the edge map image channel, after 3×3 median filtering for noise suppression.

A crucial, user-provided parameter controlling the grain of summarization is the desired number of clusters K, in clustering, and of the columns C in the summary matrix $\mathbf{C}$, in the proposed summarization method, respectively. It corresponds to the number of requested key-frames to be extracted per video. In order to most effectively compare the different algorithms and description/representation schemes, the actual number Q of different action segments (known from the ground truth) was used both as K and C for each video. The algorithm in [22] does not employ such a parameter, since summary conciseness is determined by the optimization process. Codebook size c was set to 80 for the BoF representation step. The evaluation was performed on a desktop PC with a Core i7 @ 3.5 GHz CPU and 16 GB RAM, while the codebase was developed in C++.

Table I presents the IR score averaged over the entire IMPART dataset. "Proposed" stands for the proposed algorithm, while [1] and [22] refer to the respective methods. The proposed algorithm outperforms both the established clustering technique and the sparse dictionary learning method, in terms of the IR metric.

IV. CONCLUSIONS

An algorithm for summarization of videos depicting human activities has been presented, exploiting semantic scene content properties. The proposed method extracts from an input video the video frames mostly resembling a *salient dictionary*. The dictionary component, meant to guarantee summary conciseness and compactness, identifies the video frames that only contain elementary visual building blocks. Dictionary construction is modulated by the saliency component, meant to ensure outlier inclusion and broad content coverage, which operates in the traditional inter-frame distance-based manner. Thus, the process strikes a balance between video frame representativeness and saliency. The problem is solved within an optimization framework derived from the Column Subset Selection Problem (CSSP). Empirical evaluation conducted in a publicly available dataset suggests that the presented algorithm outperforms both a baseline clustering approach and a state-of-the-art sparse dictionary learning method.

ACKNOWLEDGMENT

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

[2] 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.

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

[5] 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.

[6] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas, "Compact video description and representation for automated summarization of human activities," in *INNS Conference on Big Data*. Springer, 2016, pp. 18–28.

[7] 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.

[8] Z. Yuan, T. Lu, D. Wu, Y. Huang, and H. Yu, "Video summarization with semantic concept preservation," in *International Conference on Mobile and Ubiquitous Multimedia (MUM)*. 2011, pp. 109–112, ACM.

[9] D. P. Papadopoulos, S. A. Chatzichristofis, and N. Papamarkos, ," in *MIRAGE: Computer Vision/Computer Graphics Collaboration Techniques*. 2011, pp. 216–226, Springer.

[10] 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.

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

[12] 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.

[13] 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.

[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] 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.

[16] 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.

[17] 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.

[18] 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.

[19] G. Guan, Z. Wang, S. Lu, J. D. Deng, and D. D. Feng, "Keypoint-based keyframe selection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 729–734, 2013.

[20] S. Lu, Z. Wang, T. Mei, G. Guan, and D. D. Feng, "A Bag-of-Importance model with Locality-Constrained Coding based feature learning for video summarization," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1497–1509, 2014.

[21] 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.

[22] 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.

[23] 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.

[24] L. Duan, T. Xi, S. Cui, H. Qi, and A. C. Bovik, "A spatiotemporal weighted dissimilarity-based method for video saliency detection," *Signal Processing: Image Communicatoin*, vol. 38, no. C, pp. 45–56, 2015.

[25] 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.

[26] 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.

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