GREEDY SALIENT DICTIONARY LEARNING WITH OPTIMAL POINT RECONSTRUCTION FOR ACTIVITY VIDEO SUMMARIZATION

Ioannis Mademlis, Anastasios Tefas and Ioannis Pitas

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

ABSTRACT

Salient dictionary learning has recently proven to be effective for unsupervised activity video summarization by key-frame extraction. All relevant methods select a small subset of the original data points/video frames as dictionary atoms/representatives that, in concert, both optimally reconstruct the original entire dataset/video sequence and are salient. Therefore, they attempt to simultaneously optimize a reconstruction term, pushing towards a dictionary/summary that best reconstructs the entire dataset, and a saliency term, pushing towards a dictionary composed of salient data points. In this paper, a hypothesis is proposed and empirically tested, namely that more salient data points can be obtained by attempting to restrain reconstruction error separately for each original data point. Thus, salient dictionary learning is extended by adding a third term to the objective function, pushing towards optimal point reconstruction. A pre-existing greedy, iterative algorithm for salient dictionary learning is modified according to the proposed extension in two alternative ways. The resulting methods achieve state-of-the-art performance in three databases, verifying the validity of our hypothesis.

Index Terms— Salient dictionary learning, matrix reconstruction, video summarization, key-frame extraction

1. INTRODUCTION

Salient dictionary learning has been recently introduced for efficiently modelling the problem of activity video summarization by key-frame extraction [1] [2] [3] [4] [5]. Its purpose is to select a salient subset of representative data points from a dataset, that are both (in some sense) distinct and able to reconstruct the original dataset. Compared to traditional dictionary learning, the derived dictionary atoms are unaltered original data points, as in sparse representatives modelling [6], instead of linear combinations of original data points. Also, they are salient in the context of the entire dataset. Thus, an optimization problem is formulated that attempts to simultaneously minimize the reconstruction error induced by the dictionary and maximize the latter's saliency.

Although different uses of salient dictionary learning are possible, as variations on traditional dictionary learning applications, its efficacy has not been determined for problems other than unsupervised activity video summarization, where a dataset corresponds to an entire video sequence, each data point to a video frame and the produced dictionary to the desired video summary/key-frame set. The extracted key-frames must jointly represent the entire original video sequence as well and succintly as possible. In this domain, salient dictionary learning fits well within the methodological paradigm of previous dictionary-of-representatives approaches [6] [7] [8], which however ignored video frame saliency. The only exception known to us is [9], which considered a limited form of video frame saliency, complementary to the reconstruction term. More sophisticated, but more complex video saliency methods (e.g., [10]) have not been employed in this context. In general, for video summarization tasks, dictionary-of-representatives methods are mainly contrasted with traditional video frame/segment clustering-based approaches [11] [12] [13].

The interrelated algorithms in [1] [2] [3] [4] [5] all model the reconstruction term using the matrix Column Subset Selection Problem (CSSP) [14], i.e., the problem of optimally selecting a subset of matrix columns, so as for the subset matrix to be as close to full-rank as possible. The cardinality of the column subset, typically much smaller than the number of original matrix columns, is a fixed, pre-specified parameter. The CSSP is a NP-hard combinatorial optimization problem that can be efficiently solved only in an approximate manner. It had previously been employed for modelling movie shot selection in a summarization setting [15] [16], but it had not been used for key-frame extraction before the method in [1].

Salient dictionary learning was evaluated on summarizing activity videos, i.e., video feeds depicting human activities on a static background, with a static camera and without any editing cuts, using visual word codebook-based video frame representations [17]. Such a video, consisting in a sequence of temporally concatenated human activity segments, may come from a surveillance camera, from shooting sessions

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).

in TV/film production, etc. The ideal summary/key-frame set was implicitely defined, through the objective summarization peformance metric Independence Ratio (IR) [5], as one containing exactly one key-frame per actual activity segment (ground truth temporal activity segmentation is assumed to be known, during algorithm performance evaluation).

In general, salient dictionary learning achieved state-of-the-art results in comparison to random video frame sampling, video frame clustering [18] and two recent dictionary-of-representatives methods [8] [9]. The algorithm in [4] stood out, by achieving top performance in two out of three datasets while running in near-real-time, and highlighted the importance of the saliency term for proper video summarization. From a qualitative perspective, saliency pushes towards obtaining key-frames with more interesting visual content, instead of common but uninformative visual words (e.g., video frames simply depicting the background, or human body poses common to multiple activities). From a statistical point of view, the saliency term pushes towards inclusion of outliers which would normally be dropped by a traditional dictionary-of-representatives algorithm, since they do not contribute significantly to the reconstruction of the entire original video. Thus, by considering saliency along with reconstruction, better discrimination between different activity video segments is achieved and the Independence Ratio score is increased.

In this paper, a hypothesis is proposed and empirically tested, namely that more salient data points can be obtained by attempting to restrain reconstruction error separately for each original data point. While the reconstruction term pushes towards optimal reconstruction of the entire dataset (video) and the saliency term pushes towards a dictionary composed of salient data points (video frames), a novel third term added to the objective function will push towards optimal point reconstruction. Thus, more outlying data points will be preferred and a different form of data point saliency, implicitly derived from the optimization process, will be integrated into the framework.

A pre-existing greedy, iterative, very fast algorithm for salient dictionary learning [4] is modified according to the proposed extension in two alternative ways. Thus, two method variants are constructed: a slow one, based on the $\mathcal{L}_1$ norm of the residual error per data point, and a fast one, employing the $\mathcal{L}_2$ norm. The latter one bypasses the need to compute the residual error matrix at each iteration entirely, by taking advantage of the original algorithm properties. The resulting methods achieve state-of-the-art performance in three databases, with the fast method inducing no runtime overhead compared to [4].

2. METHOD PRELIMINARIES

Below, the general salient dictionary learning framework, the Column Subset Selection Problem and the specific salient dictionary learning algorithm from [4] are briefly reviewed.

2.1. Salient Dictionary Learning

We assume that $\mathbf{D} \in \mathbb{R}^{V \times N}$ is a representation of the original dataset/video composed of N data points/video frames, while $\mathbf{S} \in \mathbb{R}^{V \times C}$ is the desired salient dictionary/video summary/key-frame set, composed of $C << N$ unaltered columns of $\mathbf{D}$. The i-th column of $\mathbf{D}$, denoted by $\mathbf{d}_{:i}$, is the V-dimensional representation of the i-th data point/video frame. $\mathbf{s} \in \{0,1\}^N$ is a binary-valued selection vector which codifies whether each original data point will be included in the dictionary or not. $\alpha \in [0,1]$ is a user-provided parameter regulating the contribution of the saliency component. $\mathbf{p} \in \mathbb{R}^N$ is a precomputed per-point saliency vector, assigning a scalar saliency value to each original data point/video frame.

Then, the most basic, simplified form of the general salient dictionary framework [5] is the following one:

$$\min_{\mathbf{s}} :(1-\alpha)\left(\|\mathbf{D}-\mathbf{SA}\|_n\right) - \alpha(\mathbf{s}^T\mathbf{p}), \tag{1}$$

where $\|\cdot\|_n$ is a matrix norm and $\mathbf{A} \in \mathbb{R}^{C \times N}$ is a suitable coefficients matrix. Obviously, the contents of $\mathbf{S}$ depend entirely on vector $\mathbf{s}$ and the original dataset matrix $\mathbf{D}$.

2.2. The Column Subset Selection Problem

Given a matrix $\mathbf{D} \in \mathbb{R}^{V \times N}$ and a fixed value $C << N$, the matrix Column Subset Selection Problem (CSSP) consists in optimally selecting exactly C columns of $\mathbf{D}$, which jointly form a subset matrix that retains as much of the information contained in $\mathbf{D}$ as possible. Thus, the CSSP is ideal for modelling the reconstruction term in salient dictionary learning, allowing user-adjustable dictionary succinctness.

Formally, the CSSP objective is the following:

$$\min_{\mathbf{S}} : \|\mathbf{D} - (\mathbf{SS}^+)\mathbf{D}\|_F. \tag{2}$$

$\|\cdot\|_F$ is the Frobenius matrix norm and $\mathbf{S}^+$ is the pseudoinverse of $\mathbf{S}$. $\mathbf{S}$ approximates $\mathbf{D}$ in a projection sense: $\mathbf{SS}^+$ projects $\mathbf{D}$ onto the span of the C columns contained in $\mathbf{S}$.

2.3. Greedy Salient Dictionary Learning with Regularized SVD-based Saliency

By combining a regularized SVD-based method for pre-computing data point saliency vector $\mathbf{p}$ [3] and a fast, greedy, iterative method for approximately solving the CSSP [19], properly adapted to salient dictionary learning, a top-performing and near-real-time algorithm for unsupervised activity video key-frame extraction was proposed in [4]. A brief description is provided below.

Table 1: Mean IR for all competing methods across all databases (higher is better).

	$\mathcal{L}_1$**-GSD**	$\mathcal{L}_2$**-GSD**	[4]	[3]	[1]	[2]	[18]	[8]	[9]
IMPART	77.28%	$\mathbf{77.95}\%$	77.17%	72.16%	75.85	72.02%	72.94%	68.03%	50.17%
i3DPOST	$\mathbf{79.06}\%$	75.64%	77.78%	75.64%	72.56%	74.39%	72.65%	65.81%	44.87%
IXMAS	67.22%	$\mathbf{67.65}\%$	65.72%	66.38%	62.00%	66.22%	65.29%	66.16%	46.66%

Table 2: Mean runtime per video frame (in milliseconds) for all competing methods across all datasets (lower is better).

	$\mathcal{L}_1$**-GSD**	$\mathcal{L}_2$**-GSD**	[4]	[3]	[1]	[2]	[18]	[8]	[9]
IMPART	290.48	28.80	28.86	**17.90**	552.92	232.21	76.85	4043.82	427.84
i3DPOST	157.94	**31.48**	31.67	42.05	517.80	262.26	70.01	2544.20	385.35
IXMAS	593.01	49.61	**49.07**	80.82	734.34	461.15	225.45	8594.31	891.95

Initially, $\tilde{\mathbf{p}} \in \mathbb{R}^N$ is initially precomputed once. It is a slightly modified version of $\mathbf{p}$ from [3], with its entries (the per-point saliency factors) normalized into the interval $[0, 1]$. Subsequently, in the main loop, a single data point is added to the dictionary (initially empty) at each iteration, so as to greedily minimize the reconstruction error, until the key-frame set contains exactly C key-frames. The following quantities are defined for the t-th iteration:

1. $\mathbf{s}^{(t-1)}$: the currently extracted key-frame set/summary binary selection vector, prescribing the current summary $\mathbf{S}^{(t-1)}$. It holds that $\|\mathbf{s}^{(t-1)}\|_0 = t - 1$.

2. $\overline{\mathcal{R}}^{(t-1)}$: the set of the integer temporal indices of all video frames not contained in $\mathbf{S}^{(t-1)}$. It contains $N - (t-1)$ elements, all in the interval $[1, N]$.

3. $l^{(t)}$: the temporal index of the video frame $\mathbf{d}_{:l_t}$ that is actually selected for inclusion in $\mathbf{S}^{(t)}$ during iteration t. Obviously, $l^{(t)} \in \overline{\mathcal{R}}^{(t-1)}$, but $l^{(t)} \notin \overline{\mathcal{R}}^{(t)}$.

The method recursively updates two vectors, $\mathbf{f}, \mathbf{g} \in \mathbb{R}^N$. Each one keeps track of a scalar score for each video frame $\mathbf{d}_{:i}$, $0 < i \leq N$. At the start of the t-th iteration, the most suitable $l^{(t)}$ is selected for addition to the extracted key-frame set/summary in the following manner:

$$l^{(t)} = \arg\max_i \left((1-\alpha)\frac{f_i^{(t-1)}}{g_i^{(t-1)}} + \alpha \tilde{p}_i \frac{f_i^{(t-1)}}{g_i^{(t-1)}} \right), \quad i \in \overline{\mathcal{R}}^{(t-1)}. \tag{3}$$

where $f_i^{(t-1)}$, $g_i^{(t-1)}$ is the i-the entry of current vector $\mathbf{f}, \mathbf{g}$, respectively, while $\tilde{p}_i$ is the i-th entry of $\tilde{\mathbf{p}}$. Subsequently, $\mathbf{f}^{(t)}$ and $\mathbf{g}^{(t)}$ are computed according to [19], by updating $\mathbf{f}^{(t-1)}$ and $\mathbf{g}^{(t-1)}$ based on the value of $l^{(t)}$. The algorithm is completed after C iterations.

3. GREEDY SALIENT DICTIONARY LEARNING WITH OPTIMAL POINT RECONSTRUCTION

The greedy method from [4] was extended with optimal point reconstruction, so as to test our hypothesis described in Section 1. To achieve that, two alternative method variants were constructed: one based on the $\mathcal{L}_1$ norm of the residual error per data point, and one using the $\mathcal{L}_2$ norm.

In the $\mathcal{L}_1$-norm variant, we employed a recursive formula for easily updating the current residual/reconstruction error matrix $\mathbf{E}^{(t-1)} = \mathbf{D} - (\mathbf{S}\mathbf{S}^+)^{(t-1)}\mathbf{D}$ at the start of iteration t, based on $\mathbf{E}^{(t-2)}$ and $l^{(t-1)}$ [19]:

$$\mathbf{E}^{(t-1)} = \mathbf{E}^{(t-2)} - \frac{\mathbf{e}_{:l_{t-1}}\mathbf{e}_{:l_{t-1}}^T}{\mathbf{e}_{:l_{t-1}}^T\mathbf{e}_{:l_{t-1}}}\mathbf{E}^{(t-2)}, \tag{4}$$

where $\mathbf{e}_{:l_{t-1}}$ is the $l^{(t-1)}$-th column of matrix $\mathbf{E}^{(t-2)}$.

The method in [4] does not keep a residual matrix. By explicitly initializing such a matrix on the first iteration, and subsequently updating it using Eq. (4) at the start of the main loop, we obtain an estimate of the current reconstruction error (at iteration t) for the i-th data point/video frame, in the form of a per-point residual vector $\mathbf{r}^{(t-1)} \in \mathbb{R}^N$:

$$r_i^{(t-1)} = \|\mathbf{e}_{:i}^{(t-1)}\|_1, \quad 0 < i \leq N. \tag{5}$$

Subsequently, $\tilde{\mathbf{r}}^{(t-1)}$ is obtained as follows:

$$\tilde{r}_i^{(t-1)} = \frac{r_i^{(t-1)}}{\|\mathbf{r}^{(t-1)}\|_{max}}, \quad 0 < i \leq N, \tag{6}$$

in order to rescale per-point residuals into the interval $[0, 1]$. Then, we modify Eq. (3) so as to equally consider per-point saliency vector $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{r}}^{(t-1)}$ in selecting $l^{(t)}$:

$$l^{(t)} = \arg\max_i \left((1-\alpha)\frac{f_i^{(t-1)}}{g_i^{(t-1)}} + (\frac{\alpha}{2})\tilde{p}_i \frac{f_i^{(t-1)}}{g_i^{(t-1)}} + \right. \tag{7}$$

$$\left. + (\frac{\alpha}{2})\tilde{r}_i^{(t-1)} \frac{f_i^{(t-1)}}{g_i^{(t-1)}} \right), \quad i \in \overline{\mathcal{R}}^{(t-1)}.$$

By considering $\tilde{\mathbf{r}}^{(t-1)}$ at iteration t, the currently worst reconstructed data point/video frame is more likely to be selected for inclusion in the next iteration's partial dictionary/summary. If it is indeed selected, it will be optimally reconstructed from now on, both during the following iterations and, in the end, when using the finally derived dictionary. Therefore, a tendency to restrain point reconstruction error is integrated into the algorithm. Unlike $\tilde{\mathbf{p}}$, which is fixed and pre-computed, $\mathbf{r}^{(t-1)}$ is dynamically derived at each iteration from the optimization process. Thus, it constantly adapts to the current partial dictionary, until the latter's construction has been completed.

The above-described $\mathcal{L}_1$-norm greedy salient dictionary learning with optimal point reconstruction method ($\mathcal{L}_1$-GSD) is computationally inefficient, due to the need to update the reconstruction error matrix at each iteration using Eq. (4). Thus, an $\mathcal{L}_2$-norm variant of the algorithm was also developed ($\mathcal{L}_2$-GSD), taking into account the properties of the base, greedy CSSP method. Specifically, we observed that at iteration t, vector $\mathbf{g}^{(t-1)}$ encodes information about the $\mathcal{L}_2$ norm of the current residual vector per data point, due to the following relation [19]:

$$g_i^{(t-1)} = \mathbf{g}_{ii}^{(t-1)}, \quad 0 < i \leq N, \tag{8}$$

where $\mathbf{g}_{ii}^{(t-1)}$ is the i-th main diagonal entry of the symmetric matrix $\mathbf{G}^{(t-1)} = \mathbf{F}^T\mathbf{F}$, $\mathbf{F} = \mathbf{E}^{(t-1)}$. Thus, $g_i^{(t-1)} = \|\mathbf{e}_{:i}^{(t-1)}\|_2^2$, i.e., the squared $\mathcal{L}_2$ norm of the current residual vector per data point is already being implicitly computed by the algorithm, without actually needing to retain and update $\mathbf{E}$ at each iteration. From an optimization point-of-view, this can be considered a side-effect of $\mathcal{L}_2$-optimization having closed-form solutions, in contrast to $\mathcal{L}_1$-optimization.

Therefore, the $\mathcal{L}_2$-GSD method consists in replacing Eq. (5) with the following one:

$$r_i^{(t-1)} = g_i^{(t-1)}, \quad 0 < i \leq N, \tag{9}$$

and entirely discarding matrix $\mathbf{E}$. Otherwise, the algorithm is identical to the $\mathcal{L}_1$-norm variant, but performs significantly faster due to the elimination of the residual matrix update step from Eq. (4).

4. EMPIRICAL EVALUATION

The quantitative evaluation setup from [3] and [4], specially tailored for activity video key-frame extraction, was redeployed here for evaluating greedy salient dictionary learning with optimal point reconstruction. A brief description of this setup is provided below, with more relevant details available in [5].

Method comparisons were performed against a baseline clustering approach [18], as well as competing state-of-the-art methods [4] [3] [1] [2] [8] and [9]. Three specially processed activity video databases were employed, namely IMPART [20] (330 activity segments, 27252 frames at 720 × 540 pixels), i3DPOST [21] (104 activity segments, 16074 frames at 640 × 480 pixels) and IXMAS [22] (467 activity segments, 36220 frames at 390 × 290 pixels), along with the Independence Ratio (IR) objective evaluation metric.

Three different feature descriptors/modalities were extracted per video frame: LMoD [23], SIFT [24] and Improved Dense Trajectories (IDT) [25], aggregated per video frame under the Improved Fisher Vector (IFV) approach [17]. IFV codebook size was empirically set to 8, 24 and 32 visual words for IDT, SIFT and LMoD, respectively, leading to total dimensionality of video frame representation (after concatenation) $V = 17568$. In the case of [9], vectorized raw image pixel values were employed for video frame representation, due to the nature of the algorithm.

Tables 1 and 2 present the mean IR scores obtained by all competing methods, across all databases, as well as the mean execution times per video frame. Only the highest IR results across five tested values of the saliency contribution parameter (α = 0, 0.25, 0.50, 0.75, 1.00) are reported per database. Note that the corresponding IR scores of random video frame sampling averaged over a million iterations are 58.86%, 59.01% and 59.40%, for IMPART, i3DPOST and IXMAS, respectively.

As it can be seen, $\mathcal{L}_1$-GSD induces a large computational overhead compared to [4], but achieves the overall best IR score on the i3DPOST database, as well as the second best one on the IMPART and IXMAS databases. On the other hand, $\mathcal{L}_2$-GSD has practically identical runtime requirements to the very fast method in [4], while it achieves the overall best IR score on the IMPART and IXMAS databases. Thus, our hypothesis regarding the benefits of optimal point reconstruction seems to be verified, although more thorough investigation is required in the context of promising future research.

It must be noted that $\mathcal{L}_2$-GSD and [4] are, in general, the fastest methods by far. The apparent slight runtime advantage of [3] on the IMPART database is simply an artifact of presenting the evaluation results succinctly: [3] achieved its best IR performance for saliency contribution factor $\alpha = 0$, i.e., with no saliency term being computed at all.

The IR performance advantage of $\mathcal{L}_1$-GSD compared to $\mathcal{L}_2$-GSD in the i3DPOST database can be attributed to the latter containing a lower percentage of video frames that are visually outlying: most video frames are relatively similar to each other, therefore the majority are reconstructed well using a dictionary. Due to the properties of the $\mathcal{L}_2$ norm, $\mathcal{L}_2$-GSD is not able to discriminate well between video frames that have almost zero reconstruction error and video frames that have very low (but non-negligible) reconstruction error. As a result, the latter ones are incorrectly not favoured by the algorithm for inclusion in the dictionary being constructed. Unsurprisingly, the nature of the data significantly affects method behaviour.

5. CONCLUSIONS

Salient dictionary learning has been extended, using the hypothesis that integrating optimal point reconstruction into the framework would increase the saliency of the obtained dictionary. The motivation was the observation that good reconstruction of the entire dataset typically leads to increased point reconstruction error for outlier data points, which are very good candidates for conveying salient content. A very fast, top-performing greedy iterative algorithm for activity video summarization via key-frame extraction, based on salient dictionary learning, has been modified in two alternative ways, so as to incorporate optimal point reconstruction and test the proposed hypothesis. Empirical evaluation using an objective evaluation metric in three public databases indicates state-of-the-art performance and showcases the benefits of optimal point reconstruction.

6. REFERENCES

[1] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas, "Summarization of human activity videos via low-rank approximation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[2] I. Mademlis, A. Tefas, and I. Pitas, "Summarization of human activity videos using a salient dictionary," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2017.

[3] I. Mademlis, A. Tefas, and I. Pitas, "Regularized SVD-based video frame saliency for unsupervised activity video summarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[4] I. Mademlis, A. Tefas, and I. Pitas, "Greedy salient dictionary learning for activity video summarization," in *International Conference on MultiMedia Modeling (MMM) (submitted)*. 2019, Springer.

[5] I. Mademlis, A. Tefas, and I. Pitas, "A salient dictionary learning framework for activity video summarization via key-frame extraction," *Information Sciences*, vol. 432, pp. 319 – 331, 2018.

[6] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[7] 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.

[8] 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.

[9] C. Dang and H. Radha, "RPCA-KFE: Key frame extraction for video using robust principal component analysis," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3742–3753, 2015.

[10] Y. Peng, Y. Zhao, and J. Zhang, "Two-stream collaborative learning with spatial-temporal attention for video classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

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

[12] I. Mademlis, N. Nikolaidis, and I. Pitas, "Stereoscopic video description for key-frame extraction in movie summarization," in *Proceedings of the EURASIP European Signal Processing Conference (EUSIPCO)*, 2015.

[13] M. Otani, Y. Nakashima, E. Rahtu, J. Heikkilä, and N. Yokoya, "Video summarization using deep semantic features," *arXiv preprint arXiv:1609.08758*, 2016.

[14] C. Boutsidis, M. W. Mahoney, and P. Drineas, "An improved approximation algorithm for the Column Subset Selection Problem," in *Proceedings of the Symposium on Discrete Algorithms*, 2009.

[15] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas, "Movie shot selection preserving narrative properties," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2016.

[16] I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas, "Multimodal stereoscopic movie summarization conforming to narrative characteristics," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5828–5840, 2016.

[17] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher Kernel for large-scale image classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2010.

[18] 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.

[19] A. K Farahat, A. Ghodsi, and M. S. Kamel, "Efficient greedy feature selection for unsupervised learning," *Knowledge and Information Systems*, vol. 35, no. 2, pp. 285–310, 2013.

[20] T. Theodoridis, A. Tefas, and I. Pitas, "Multi-view semantic temporal video segmentation," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2016.

[21] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3DPOST multi-view and 3D human action/interaction database," in *Proceedings of the IET Conference for Visual Media Production (CVMP)*, 2009.

[22] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 249–257, 2006.

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

[24] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1999.

[25] H. Wang and C. Schmid, "Action recognition with Improved Trajectories," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013.