REGULARIZED SVD-BASED VIDEO FRAME SALIENCY FOR UNSUPERVISED ACTIVITY VIDEO SUMMARIZATION

Ioannis Mademlis, Anastasios Tefas and Ioannis Pitas

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

ABSTRACT

Storage, browsing and analysis of human activity videos can be significantly facilitated by automated video summarization. Unsupervised key-frame extraction remains the most widely applicable technique for summarizing activity videos. However, their specific properties make the problem difficult to solve. Typical relevant algorithms fall under the video frame clustering or the dictionary-of-representatives families, with salient dictionary learning having been recently proposed. Under this formulation, the video frames selected as key-frames are the ones which simultaneously best reconstruct the entire video and are salient compared to the rest. This paper improves upon such a method by replacing the video frame saliency estimation term with one based on Regularized SVD-based Low Rank Approximation, taking advantage of the well-established correlation between mid-range matrix singular values and salient regions. Extensive empirical evaluation showcases the high performance of both the salient dictionary learning framework and the specific proposed method.

Index Terms— Video Summarization, Big Data, Video Saliency, Singular Value Decomposition, Key-frame Extraction

1. INTRODUCTION

Human activity videos constitute a common target for automated video summarization algorithms, since they typically extend to many hours of mostly uninteresting footage, while only a small percentage of the video frames are actually important. Such videos (derived from surveillance feeds, TV/film production shooting sessions, sports coverage with static camera, etc.) have certain recurrent characteristics: heavy inter-frame visual redundancy due to static camera and static background, lack of editing cuts, lack of objectivity in specifying the most important video frames. Video summarization is a difficult task where a delicate balance between different factors has to be achieved in the produced summary, including sufficient compactness (lack of redundancy), conciseness, outlier inclusion, semantic representativeness and content coverage. The specific properties of activity videos further complicate the problem.

Different forms of activity video summaries have emerged over the years. The most trivial one is simple *temporal video segmentation* [1], where the video has to be partitioned over time in consecutive activity segments. This is actually a substitute of shot cut/boundary detection, commonly applied in other types of video summarization as a pre-processing step [2]. Therefore, a second summarization step is still needed. Another approach is *video synopsis* [3], where the summary consists in a number of synthetic video frames derived from blending visually active regions found in multiple original video frames. This cannot be applied if the goal is to retain a subset of the original video frames without processing them (e.g., in sports coverage), or if the depicted scenes are crowded and contain overlapping active regions. Finally, *skimming* [4] results in a short "trailer", but requires specific video frames to be pre-identified as important. Thus, *key-frame extraction* is the most significant stage of a video summarization pipeline. A *static video summary* is composed of a set of temporally ordered "key-frames", i.e., a subset of the original video frame set.

Unsupervised methods mainly fall under two categories: video frame clustering and dictionary-of-representatives approaches. In the first case, the video frames are partitioned in distinct groups and the ones closest to the cluster centroids are selected as the key-frames [5] [6]. The number of clusters/key-frames either depends proportionally on the video length, or is defined by the user. Such distance-based data partitioning approaches do not take into account the semantic content of video frames, completely offloading semantics to the underlying video frame representation method. On the other hand, dictionary-of-representatives-based algorithms assume that all video frames can be reconstructed as linear combinations of a small number of representatives among them, which are subsequently identified and selected as the key-frames [7] [8]. In most cases, the cardinality of the representatives/key-frame set is pre-fixed by the user. These approaches inherently consider video frame semantics in an unsupervised manner, since they detect video frames containing isolated visual building blocks of the original video. The video frame representations upon which these methods operate are vectors derived using global image descriptors [9]

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).

[6], local image descriptors aggregated under a representation model (such as Bag-of-Features) [10] [11], or the raw image pixel values [8].

Recently, salient dictionary learning was introduced for activity video key-frame extraction [12]. Under this formulation, the key-frame set is extracted by simultaneously optimizing the desired summary for maximum reconstructive ability and maximum saliency. The reconstruction term guarantees summary conciseness, representativeness and compactness, inherently operating within the constraints imposed by video semantics, through identification of the video frames containing only the elementary visual building blocks. Dictionary construction is modulated by the saliency term, meant to ensure outlier inclusion and broad content coverage, which is computed in a simple inter-frame distance-based manner. The entire process is expressed as a Column Subset Selection Problem (CSSP) [13]. A specific form of video frame saliency was also integrated into generic video summarization in [8], although in a much more inflexible manner.

The numerical, Singular Value Decomposition-based algorithm in [12] improved upon a previous, slower, no-saliency CSSP-based genetic approach that was employed for activity video summarization in [14]. Both methods extract a pre-defined number C of key-frames. Their performance was empirically measured using a novel objective evaluation metric that bypassed the subjective, or semi-subjective, nature of traditional summarization metrics. Temporal video segmentation ground-truth annotation data, describing obvious temporal boundaries between consecutive activity video segments, were employed for counting the number of extracted key-frames derived from actually different activity segments (*independent key-frames*), as an indirect indication of summarization success. For evaluation purposes, the total number of requested key-frames per video (C) was set equal to the corresponding real number of different activity segments (known from the ground truth). Thus, the ratio of extracted independent key-frames to C, called *Independence Ratio* (IR) score, was employed as a practically objective evaluation metric, with any two video frames belonging to the same activity segment treated as interchangeable.

Despite the high performance (in terms of IR) and speed of the algorithm presented in [12], the simple per-frame saliency computation is particularly time consuming (since a dense inter-frame distance matrix must be constructed) and, by design, only takes into account temporally local saliency, i.e., the saliency of each video frame mainly depends on its distance from its temporal neighbours. This paper addresses the above limitations by replacing the saliency term with one based on the SVD decomposition of the original video frame matrix. The SVD decomposition is readily available, since it is necessary for computing the reconstruction term and, therefore, the proposed saliency term only adds minimal computation overhead. Additionally, the proposed term takes a global perspective while evaluating per-frame saliency, by exploiting the well-established correlation between mid-range matrix singular values and salient regions.

2. ALGORITHM BACKGROUND

Below, 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 $\mathbf{C} \in \mathbb{R}^{V \times C}$, $C << N_f$ containing an ordered set of video key-frames. Its columns are indicated by a binary-valued frame selection vector $\mathbf{s} \in \mathbb{N}^{N_f}$.

2.1. The Column Subset Selection Problem

In the methods this paper improves upon (the no-saliency, dictionary-of-representatives algorithm [14] and the salient dictionary learning algorithm [12]), the Column Subset Selection Problem (CSSP) [13] was selected for algebraically modelling the reconstruction term.

Given $\mathbf{D}$ and a parameter $C << N_f$, the 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 $\mathbf{C}$.

The CSSP is considered to be NP-hard and, besides exhaustive search, only approximate solutions are known. A fast, numerical, randomized method operating in two stages [13] was employed as a main building block in [12]. First, approximately $C\log C$ columns are sampled from matrix $\mathbf{D}$, using an SVD-derived probability distribution. Thus, the majority of the less outlying columns are removed and a preliminary summary matrix is obtained that serves as a suitable input to the second stage. Then, exactly C columns can be deterministically selected from the sample using any traditional CSSP algorithm (the Rank-Revealing QR decomposition [15] was employed in [12]).

2.2. Salient Dictionary Learning for Activity Summarization

Human activity videos are mainly composed of elementary visual building blocks assembled in several combinations, thus $\mathbf{D}$ is assumed to be low-rank. A salient dictionary learning algorithm consists in the simultaneous optimization of two components: the "reconstruction term" and the "saliency term". Intuitively, the reconstruction term alone will tend to favour video frames solely containing common, elementary visual building blocks of the entire video, which facilitate the reconstruction process. This leads to inclusion of uninteresting video frames (e.g., depicting the static background)

and exclusion of outliers from the summary. A saliency term becomes necessary to compensate for this, leading to the so-called *salient dictionary learning* objective defined in [12]:

$$\min_{\mathbf{s}} : \|\mathbf{D} - \mathbf{C}\mathbf{C}^{+}\mathbf{D}\|_F - \alpha c \mathbf{s}^T \mathbf{p}, \quad (2)$$

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. $\mathbf{p} \in \mathbb{R}^{N_f}$ is a precomputed per-frame saliency vector, assigning a scalar saliency value to each video frame.

In [12], the approximate CSSP algorithm from [13] was coupled with a simple saliency term, in order to solve the salient dictionary learning problem for activity video summarization. The saliency term was adapted from the spatial, intra-frame component of the saliency estimation algorithm presented in [16]. By exploiting a dense inter-frame distance matrix, saliency values were assigned to entire video frames, instead of video frame blocks, and spatial distance between the latter ones was replaced by temporal distance between video frames. The method resulted in a pre-computed, per-frame saliency vector $\mathbf{p}$.

Subsequently, in order to adapt the employed CSSP method from [13] to salient dictionary learning, matrix $\mathbf{D}$ was modified in the following manner:

$$\hat{\mathbf{D}} = (1 - \alpha)\mathbf{D} + \alpha \mathbf{D} \left(\mathbf{diag}(\mathbf{n})\mathbf{diag}(\mathbf{p})\right), \quad (3)$$

where $\mathbf{n} \in \mathbb{R}^{N_f}$ was 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) were scaled down to a degree directly proportional to their saliency and to the provided saliency contribution parameter α. Finally, the numerical algorithm in [13] was applied on $\hat{\mathbf{D}}$, in order to obtain the desired summary. The above method implicitly solved the objective from Eq. (2).

3. REGULARIZED SVD-BASED LOW RANK APPROXIMATION FOR SALIENT DICTIONARY LEARNING

Despite the good performance and fast execution of the salient dictionary learning method in [12], compared to simple clustering and to the no-saliency, genetic CSSP algorithm from [14], the relatively high time requirements of constructing the dense inter-frame distance matrix for computing the simple saliency term, as well as the temporally local nature of the estimated video frame saliency, were clear limitations of the algorithm.

To mitigate this issue, the correlation between mid-range matrix singular values and salient regions (well-established in image saliency estimation [17]) is exploited here in order to formulate an alternative saliency term. The proposed method models precomputed per-frame saliency on a regularized SVD-based reconstruction of $\mathbf{D}$. The SVD decomposition is already employed for computing the reconstruction term (by the CSSP algorithm [13]) and, therefore, the additional computational overhead introduced by the proposed saliency term is minimal. This is in contrast to the saliency term used in [12], where the time needed for computing the inter-frame distance matrix was significant.

First, the SVD decomposition $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is obtained. Then, the singular values of $\mathbf{D}$, lying ordered on the diagonal of $\mathbf{\Sigma}$, are clustered into three groups: large, intermediate and small. To achieve this, the singular values are adaptively clustered into three discrete groups (large, intermediate, small) using a fast, dynamic programming-based variant [18] of the Jenk's Natural Breaks Optimization algorithm for one-dimensional clustering [19]. The latter operates by exploiting a scalar version of the Fisher ratio, typically employed in Linear Discriminant Analysis (LDA), thus by attempting to simultaneously minimize intra-cluster variance and maximize inter-cluster variance. The large ones and the small ones among the singular values are set to zero and, thus, the regularized matrix $\tilde{\mathbf{\Sigma}}$ is derived. Subsequently, the video matrix is approximately reconstructed using $\tilde{\mathbf{\Sigma}}$:

$$\tilde{\mathbf{D}} = \mathbf{U}\tilde{\mathbf{\Sigma}}\mathbf{V}^T. \quad (4)$$

In image saliency estimation, the underlying intuition would be that large, intermediate and small singular values correspond to non-salient/visually dominating image regions (e.g., the background), salient/important image regions and noise/fine-grained visual details, respectively. In the proposed method, the video frame representation $\mathbf{D}$ (encoding spatiotemporally varying content) is employed in place of raw image data (directly conveying spatially varying content). Thus, in $\tilde{\mathbf{D}}$, salient spatiotemporal video regions have been enhanced and noise or non-salient regions have been suppressed. $\tilde{\mathbf{D}}$ is, in essence, a two-dimensional spatiotemporal video saliency map.

A preliminary saliency value for the i-th video frame can easily be extracted from $\tilde{\mathbf{D}}$ in the following manner:

$$\tilde{\mathbf{p}}_i = \|\tilde{\mathbf{d}}_{:i}\|_1, \quad (5)$$

where $\tilde{\mathbf{d}}_{:i}$ is the the i-th column of $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{p}}$ is a preliminary, per-frame saliency vector.

The final, precomputed per-frame saliency vector $\mathbf{p}$ can then be derived by applying the following post-processing saliency enhancement step on $\tilde{\mathbf{p}}$. Initially, the preliminary saliency value $\tilde{\mathbf{p}}_i$ of video frame $\mathbf{d}_{:i}$ is subtracted from the average saliency of its temporal neighborhood $[i-M, i+M]$. This is implemented by first performing moving average filtering on $\tilde{\mathbf{p}}$, using a filtering window of length $2M+1$. Subsequently, all negative per-frame saliency values (corresponding to video frames which, on average, are less salient than their neighbours) are set to zero, giving rise to the final precomputed, per-frame saliency vector $\mathbf{p}$.

Table 1. Mean IR scores for all competing methods across all datasets (higher is better).

	Random	**Proposed**	[14]	[12]	[9]	[7]	[8]
IMPART	58.86%	72.16%	**75.85%**	72.02%	72.94%	68.03%	50.17%
i3DPOST	59.01%	**75.64%**	72.56%	74.39%	72.65%	65.81%	44.87%
IXMAS	59.40%	**66.38%**	62.00%	66.22%	65.29%	66.16%	46.66%

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

	Proposed	[14]	[12]	[9]	[7]	[8]
IMPART	**17.90**	552.92	232.21	76.85	4043.82	427.84
i3DPOST	**42.05**	517.80	262.26	70.01	2544.20	385.35
IXMAS	**80.82**	734.34	461.15	225.45	8594.31	891.95

The intuition behind this post-processing step is that the most salient video frames should be temporally distant, similarly to how salient image regions are typically selected so as to be spatially distant, with less salient regions suppressed, in image saliency map estimation algorithms [17]. Such a consideration also fits well with video summarization, where the demand of maximum content coverage requires the extracted key-frames to be temporally dispersed.

As soon as the final $\mathbf{p}$ has been computed, $\mathbf{D}$ can be modified with Eq. (3). Then, the numerical algorithm from [13] can be applied on $\hat{\mathbf{D}}$ in order to obtain the desired summary, as in [12].

4. QUANTITATIVE EVALUATION

In order to empirically evaluate the proposed algorithm, extensive comparisons were made against a baseline clustering approach [9], random video frame sampling over a million iterations, as well as competing state-of-the-art methods [14] [12] [7] and [8], using three human activity video datasets. This also serves as a much more comprehensive empirical evaluation for the methods in [14] and [12] than the single-dataset evaluation results available there.

A richer video frame description and representation scheme was now employed, compared to [14] and [12]. Three different feature descriptors/modalities were extracted per video frame: LMoD [11], SIFT [20] and IDT [21], aggregated per video frame under the IFV approach [22]. IFV codebook size was empirically set to 8, 24 and 32 visual words for IT, SIFT and LMoD, respectively, leading to total dimensionality $V = 17568$. This description/representation scheme was selected due to its consistent performance across datasets. The much weaker and highly dataset-tuned (w.r.t parameters) scheme employed in [14] and [12] led, for all summarization methods, to very high performance on the single dataset employed there, but too low IR scores on other datasets. In the case of [8], vectorized raw image pixel values were employed for video frame representation, due to the nature of the algorithm.

Single-view subsets of three publicly available, annotated activity video datasets were employed. The datasets were slightly processed (e.g., multiple consecutive activities concatenated) to suit the task. The datasets are IMPART [1] (330 activity segments, 27252 frames at 720 $\times$ 540 px), IXMAS [23] (467 activity segments, 36220 frames at 390 $\times$ 290 px) and i3DPOST [24] (104 activity segments, 16074 frames at 640 $\times$ 480 px).

Tables 1 and 2 present the mean IR scores obtained by all competing methods, across all datasets, as well as the mean execution times per video frame. The significant differences in evaluation results from the ones in [14] and [12] have arisen due to the different representation scheme employed here, the removal of time needed for video frame description/representation from the reported runtime, as well as the fact that the mean results over five iterations are shown here for non-deterministic algorithms, instead of the best results. Additionally, for [12], [8] and the proposed method, only the highest IR results across five tested values of the saliency contribution parameter ($\alpha = 0$, 0.25, 0.50, 0.75, 1.00) are reported per dataset.

As it can be seen, [8] completely fails to handle activity summarization, simple clustering from [9] performs surprisingly well, while [14] achieves best IR performance in IMPART, at a significant runtime penalty compared to the proposed method which is, by far, the fastest of all. Additionally, the proposed algorithm achieves top IR performance in the other two datasets.

5. CONCLUSIONS

A method for activity video key-frame extraction via salient dictionary learning has been presented. The proposed approach entails replacing the simple saliency term from an earlier algorithm with a much faster to compute term, that inherently considers video frame saliency from a global perspective. The proposed saliency term is based on Singular Value Decomposition, thus fitting very well with the reconstruction term from the earlier algorithm (an SVD-based solution to the Column Subset Selection Problem). Extensive empirical evaluation conducted on three processed public datasets suggests that, in general, the presented algorithm achieves state-of-the-art performance at near-real-time execution time.

6. REFERENCES

[1] 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.

[2] 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.

[3] X. Song, L. Sun, J. Lei, D. Tao, G. Yuan, and M. Song, "Event-based large scale surveillance video summarization," *Neurocomputing*, vol. 187, pp. 66–74, 2016.

[4] 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.

[5] 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.

[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] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] T. F. Chan and P. C. Hansen, "Low-rank revealing QR factorizations," *Numerical Linear Algebra with Applications*, vol. 1, no. 1, pp. 33–44, 1994.

[16] 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.

[17] X. Ma, X. Xie, K.-M. Lam, J. Hu, and Y. Zhong, "Saliency detection based on Singular Value Decomposition," *Journal of Visual Communication and Image Representation*, vol. 32, pp. 95–106, 2015.

[18] M. Hilferink, "Fisher's Natural Breaks classification," http://wiki.objectvision.nl/index.php/Fisher2013.

[19] G. F. Jenks, "The data model concept in statistical mapping," *International Yearbook of Cartography*, vol. 7, no. 1, pp. 186–190, 1967.

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

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

[22] 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, pp. 143–156.

[23] 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.

[24] 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 IEEE Conference for Visual Media Production (CVMP)*, 2009, pp. 159–168.