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

Ioannis Mademlis, Anastasios Tefas, Ioannis Pitas

*Department of Informatics,
Aristotle University of
Thessaloniki*

Background

- Video summarization: generating condensed versions of a video, through the identification of its most important and salient content.
- The abstracted content to be included in the target summary can be represented as a carefully selected subset of the original video frames, i.e., a key-frame set.
- Different needs must be balanced when deriving the summary: representativeness / content coverage, outlier inclusion, compactness (lack of redundancy) and conciseness.
- Activity videos summarization is a special case with wide applicability (e.g., surveillance feeds, sports footage, film/TV production). Its properties (static camera, static background, lack of clearly discernible shot cut/boundaries) require special handling.

Summary

- This work modifies a recently proposed salient dictionary learning algorithm for activity video key-frame extraction [2]: the produced key-frame set consists in the original video frames that are most able to linearly reconstruct the entire video content (*reconstruction term*) and, simultaneously, most salient *(saliency term)*.
- In [2] the problem is cast as a matrix Column Subset Selection Problem (CSSP) and approximately solved by a numerical, SVD-based algorithm for the CSSP, properly adapted so as to take into account a saliency score per video frame.
- The simple per-frame saliency measure in [2] is particularly time-consuming (a dense inter-frame distance matrix is required) and only considers local saliency (the saliency of each video frame mainly depends on its distance from its temporal neighbors).
- This paper replaces the saliency measure with one based on the SVD decomposition of the original video data matrix, which is readily available due to the employed reconstruction term algorithm.
- Moreover, the proposed saliency 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.

Column Subset Selection Problem (CCSP)

- The CSSP is a combinatorial optimization problem, considered to be NP-hard.
- By employing CSSP, and contrary to standard dictionary learning, the learnt dictionary atoms consist in unaltered, original data points.

–$M \times N$ matrix **D**, parameter $C << N$

–Goal: select a subset of exactly C columns of **D**, to form a new $M \times C$ matrix **C** that approximates **D**, while being as close to full-rank as possible

–Minimize: $\|\mathbf{D} - (\mathbf{C}\mathbf{C}^{+})\mathbf{D}\|_F$

–$\|\cdot\|_F$ is the Frobenius matrix norm and $\mathbf{C}^{+}$ is the pseudoinverse of **C**.

Activity Video Summarization Based on Salient Dictionary Learning

- Each video frame is described and represented as a single vector, using an Improved Fisher Vector (IFV) aggregation approach.
- The CSSP is employed as a reconstruction term
- The complete salient dictionary learning objective is the following one:

$$\min_{\mathbf{s}} : \|\mathbf{D} - \mathbf{C}\mathbf{C}^{+}\mathbf{D}\|_F - \alpha c \mathbf{s}^T \mathbf{p}$$

- Notations: N_f is the total number of original video frames,

–V is the dimensionality of each video frame representation

–**s** is a N_f-dimensional binary video frame selection vector

–**p** is a N_f-dimensional video frame pre-computed saliency score vector

–α is a user-provided saliency term contribution weight

–c is a scaling factor bringing per-video frame saliency value down to the scale of the reconstruction term

–C is the desired extracted key-frame set cardinality

–**D** is the $V\times N_f$ original data matrix (video frame set)

–**C** is the desired $V \times C$ summary (key-frame set), constructed using **s**

- The goal is to find the matrix **C**, with its columns being unaltered columns of **D**, that minimizes the objective.
- In [2], an approximate SVD-based, two stage CSSP algorithm [1] is adopted for solving the problem.
- Before applying the CSSP algorithm, matrix **D** is properly modified in order to take into account a per-video frame saliency measure. In the modified matrix, less salient columns have been scaled down in norm to a degree directly proportional to their saliency and a user-provided saliency term contribution weight.

Regularized SVD-based video frame saliency

- [2] is modified here by replacing the simple saliency measure (employed for precomputing **p**) with a faster, SVD-based approach. Since the SVD decomposition $\mathbf{D}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is already used for the evaluating the reconstruction term (in the CSSP algorithm), the computational overhead of this saliency measure is minimal.
- First, the singular values of **D**, lying ordered on the diagonal of **Σ**, are clustered into three groups: large, intermediate and small. This is achieved using a fast variant of the Jenk's Natural Breaks Optimization method for one-dimensional clustering, that operates by exploiting a scalar version of the Fisher Ratio.
- The large and the small singular values are set to zero. Thus the regularized matrix $\tilde{\mathbf{\Sigma}}$ is derived.
- Then, the video matrix is approximately reconstructed: $\tilde{\mathbf{D}} = \mathbf{U}\tilde{\mathbf{\Sigma}}\mathbf{V}^T$.

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).

Regularized SVD-based video frame saliency

- 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 **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: $\tilde{\mathbf{p}}_i = \|\tilde{\mathbf{d}}_{:i}\|_1$

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-video frame saliency vector **p** can then be derived from $\tilde{\mathbf{p}}$ by applying a simple, saliency enhancement step based on moving average filtering. 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.
- Subsequently, the method from [2] may be employed for salient dictionary learning.

Evaluation

- Extensive comparisons were made against a baseline clustering approach [3], random video frame sampling over a million iterations, as well as competing state-of-the-art methods [4] [2] [5] and [6], using three human activity video datasets (IMPART, i3DPOST, IXMAS). The objective IR (Independence Ratio) activity summarization metric has been adopted from [4] and [2].

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

	Random	**Proposed**	[4]	[2]	[3]	[5]	[6]
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	[4]	[2]	[3]	[5]	[6]
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

References

[1] 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.

[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] 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.

[4] 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.

[5] 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.

[6] 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.