ICIP 2017

Summarization of Human Activity Videos Using a Salient Dictionary

Ioannis Mademlis
Anastasios Tefas
Ioannis Pitas

Department of Informatics,
Aristotle University of Thessaloniki, Greece

Department of Electrical and Electronic Engineering,
University of Bristol, UK

Video Summarization

- Problem: Applications producing very large volumes of video data on a daily basis (e.g., surveillance streams, professional capture sessions in media production)

- Solution: In several scenarios, a properly constructed summary of the video feed may be employed instead of the lengthier original video, thus significantly reducing storage and processing requirements

- *Video summarization* aims at generating such condensed versions of a video stream, by only selecting and retaining its most important and representative content

- The abstracted content can be represented/stored as a temporally ordered sequence of video frames (key-frames)

IMP-ART

AIIA Laboratory
Artificial Intelligence Information Analysis

Video Summarization

- Several approaches to summarization have been developed over the years:
  - **Clustering**
  - **Graph Optimization**
  - **Sparse Dictionary Learning**
  - …
- The original video is typically represented as a set of vectors, with each one corresponding to a video frame. Three approaches:
  - Local feature-based aggregation methods (BoF, FV, VLAD)
  - Global frame features
  - Spatial subsampling $\longrightarrow$ raw pixel values $\longrightarrow$ vectorization

Summarization of Human Activity Videos

- Task: Summarize video sequences depicting human subjects performing actions
  - Ideally, automatically extract one key-frame per depicted action
  - Static camera
  - Static background
  - No shot cut boundaries available (one long take)
  - Heavy intra-frame visual content redundancy

Summarization of Human Activity Videos

- <u>Idea</u>: Select video frames which simultaneously best reconstruct the entire video sequence and are salient
  - Optimization problem with a reconstruction term and a saliency term
  - The reconstruction term reflects the fact that human activity videos are mainly composed of elementary visual building blocks composed in various combinations
  - The saliency term is necessary to avoid common but uninteresting video frames (e.g., depicting only the background)
  - Thus, the optimization must converge to a *salient dictionary* of unaltered, original video frames
- <u>Video representation:</u> *V* x *N* matrix **D**, where *N* is the number of video frames

Column Subset Selection Problem (CSSP)

- We modeled the reconstruction term as a Column Subset Selection Problem (CSSP):
  - NP-hard combinatorial optimization problem
  - Given a *V* x *N* matrix **D**, the goal is to select *C* << *N* columns, so that the *V* x *C* matrix **C** (composed of the selected columns) captures as much information contained in **D** as possible:

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

  - $\mathbf{CC}^{+}$ is a projection matrix onto the span of the columns of **C**
  - Minimizing (1) means finding a subset matrix **C** that is as close to full-rank as possible

Local Outlier Detection

- We modeled the saliency term as a simple, local outlier detection problem:
  - A scalar saliency value is pre-computed for each column/video frame
  - The result is a per-frame, N-dimensional saliency vector $\mathbf{p}$:

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

where $\mathbf{d}_i$ is the i-th column of $\mathbf{D}$, i.e., the i-th video frame

  - The more different a video frame is to its temporal neighbors, the more salient it is considered to be

The optimization problem

- Implicitly, we optimize the following criterion:

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

where:

- $\alpha \in [0, 1]$ is a user-provided parameter regulating the contribution of the saliency component
- c is an optional scaling factor
- $\mathbf{s} \in \{0, 1\}^N$ is the desired solution: a binary-valued video frame selection vector
- **C** is the *V* x *C* summary matrix constructed based on **s** ($\|\mathbf{s}\|_1 = C$)

IMPART

Problem Solution

- We extend a landmark SVD-based method for solving the CSSP [Boutsidis2009]
- The method includes a randomized first stage and a deterministic second one:
  - First, approximately $C\log C$ columns are randomly sampled from **D**
  - Sampling follows a distribution p_i constructed using the top-C right singular subspace of **D**, spanned by the columns of the SVD-provided N x C matrix $\mathbf{V}_C$:

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

    where p_i is the probability of selecting the i-th column of **D** and $(\mathbf{V}_C)_i$ is the i-th row of $\mathbf{V}_C$
  - In the second stage, exactly C columns are selected from the sample using a deterministic algorithm

Problem Solution

- Intuition: the employed sampling distribution defined over the original matrix columns is actually the normalized statistical leverage scores of the columns
- Thus, a preliminary summary is initially constructed containing the more globally outlying columns
- Subsequently, we employ a traditional deterministic method used for CSSP (RRQR), in order to select exactly *C* columns from the longer preliminary summary

Problem Solution

- Modification: In order to adapt the above method to our proposed approach, we pre-modify video matrix **D** according to the per-frame saliency vector **p**:

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

where **n** contains normalization coefficients mapping saliency factors to the interval [0, 1]

- In $\hat{\mathbf{D}}$ less salient columns/video frames have been scaled down, to a degree directly proportional to their saliency and to the contribution parameter α
- Subsequently, the two-stage CSSP algorithm is applied on $\hat{\mathbf{D}}$

IMP-ART

Frame Representation

- Local Moments Descriptor [Mademlis2016]
  - Local image descriptor capturing low-level image statistics from several image channels (luminance, color hue, optical flow magnitude, edge map)
- Trajectories component from Dense Trajectories [Wang2013]
  - Spatiotemporal activity descriptor conveying semantic content
- Bag-of-Features feature aggregation: each video frame is represented as a distribution of elementary visual building blocks

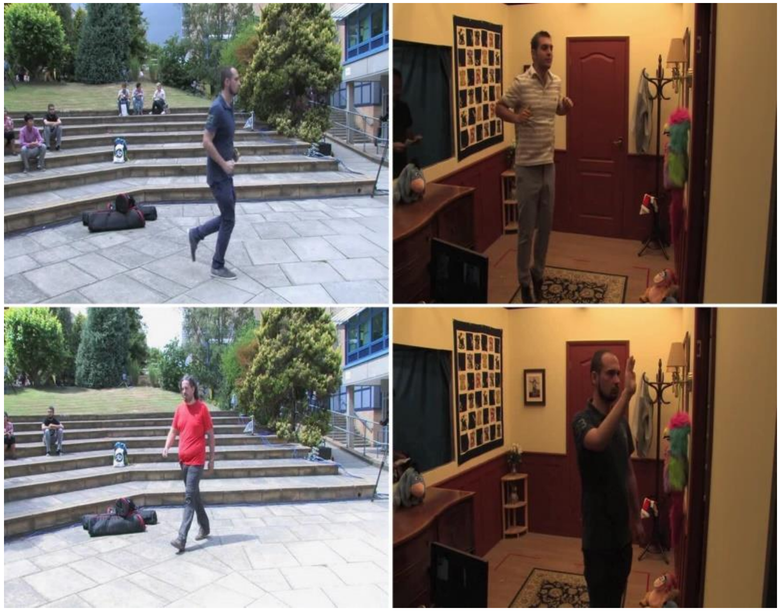
IMPART

Artificial Intelligence Information Analysis Laboratory

Evaluation

- **IMPART** dataset
- 6 720x540 videos, where each one depicts three actors performing a series of consecutive actions, with a static camera and static background
- The number of extracted key-frames derived from actually different activity segments (hereafter called *independent key-frames*) can be used as an objective indication of summarization success (as in [Mademlis2017])
- **Independence Ratio** metric: Independent / Total Key-Frames
- α (saliency contribution) was set to 0.25
- C (number of key-frames) was set to the actual number of ground-truth activity segments

Evaluation Results



- Comparison with:
  - A baseline clustering approach [Avilla2011]
  - A sparse dictionary learning algorithm for incremental video summary construction [Mei2015]
  - An early version of our method that does not consider saliency and solves the CSSP via a genetic algorithm [Mademlis2017]

	Proposed	[Avilla2011]	[Mei2015]	[Mademlis2017]
IR	**0.872**	0.571	0.802	0.840
Msecs/frame	**33.65**	54.78	1113.56	409.32

IMP-ART

Conclusions

- The proposed definition of a video summary as a salient dictionary of key-frames seems to suit well the activity video summarization task
- The specific proposed algorithm is both faster and more accurate than the tested competing approaches
- Further work needs to be done towards exploring:
  - Different types of reconstruction and saliency terms
  - Summarization of different video types
  - Better adjustment of the frame representation to the summarization task (e.g., via learnt features)

References

- [Avilla2011]: 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
- [Boutsidis2009]: 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, Society for Industrial and Applied Mathematics
- [Mademlis2016]: 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. 2016, Springer
- [Mademlis2017]: 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
- [Mei2015]: 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
- [Wang13]: H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "*Dense trajectories and motion boundary descriptors for action recognition*," International Journal of Computer Vision, vol. 103, no. 1, pp. 60–79, 2013

Thank you for your attention!