SUMMARIZATION OF HUMAN ACTIVITY VIDEOS VIA LOW-RANK APPROXIMATION

Ioannis Mademlis, Anastasios Tefas, Nikos Nikolaidis, Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki

Department of Electrical and Electronic Engineering University of Bristol

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)

–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 handle.

Summary

•This work presents a key-frame extraction algorithm for activity videos that operates by selecting the subset of the original video frames most able to linearly reconstruct the entire video content in an accurate manner.

•Such an approach can be included in a recent wave of video summarization methods, based on learning a dictionary of representative video frames.

•To ensure conciseness, the cardinality C of the key-frame set is pre-fixed and user-defined. This can be viewed as an advantage over competing methods, where conciseness is only enforced via a sparsity constraint during optimization.

•The problem is cast as a matrix Column Subset Selection Problem (CSSP) and solved by a genetic algorithm, without resorting to convex relaxation.

•Until now, the CSSP has mainly been exploited for feature selection. It has not been employed before for key-frame extraction.

Column Subset Selection Problem (CCSP)

•The CSSP is an NP-hard and non-convex combinatorial problem, related to sparse dictionary learning and low-rank approximation.

•Contrary to standard sparse dictionary learning, the learnt dictionary atoms consist in unaltered, original data points.

–$M \times N$ matrix $\mathbf{D}$, parameter $C << N$

–Goal: select a subset of exactly C columns of $\mathbf{D}$, to form a new $M \times C$ matrix $\mathbf{C}$ that approximates $\mathbf{D}$, 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 $\mathbf{C}$.

Activity Video Summarization Based on the CSSP

•Each video frame is described and represented as a relatively low-dimensional vector, using an image descriptor and a Bag-of-Features (BoF) aggregation approach.

•Several image channels are independently described per video frame and the corresponding representation vectors are concatenated.

•No knowledge/existence of shot cuts/boundaries is required.

•A reasonable assumption is made, i.e., activity videos are approximately composed of elementary visual building blocks, assembled in several different linear combinations.

•Notations:

– N_f is the total number of original video frames is,

– c is the BoF codebook size per image channel

– K is the number of image channels per video

– C is the desired extracted key-frame set cardinality

– $\mathbf{D}$ is the $Kc \times N_f$ original data matrix (video frame set)

– $\mathbf{C}$ is the desired $Kc \times C$ summary (key-frame set)

•The goal is to find the matrix $\mathbf{C}$, with its columns being unaltered columns of $\mathbf{D}$, that minimizes the CSSP objective.

•$\mathbf{C}\mathbf{C}^{+}\mathbf{D}$ is the low-rank projection of $\mathbf{D}$ onto the span of the columns of $\mathbf{C}$.

Genetic Solution to the CSSP for Activity Video Summarization

•The desired solution is a set of matrix indices with pre-fixed cardinality C. Since $\mathbf{D}$ is a $Kc \times N_f$ matrix, for the k-th index with an assigned value g_k the following hold:

$$k \in \mathbb{N}, \quad k \in [1, \cdots, C].$$

$$g_k \in \mathbb{N}, \quad g_k \in [1, \cdots, N_f].$$

•Each candidate/chromosome is encoded in the form of a sequence of column indices.

•Roulette selection.

•Genetic operators: order preserving variants of 1-point crossover and mutation [1].

•Fitness function: $f(\mathbf{h}_n) = \| \mathbf{D} - \mathbf{C}_n\mathbf{C}_n^{+} \mathbf{D} \|_F^{-1}$, where $\mathbf{h}_n$ is the n-th candidate in the current population.

•Pre-fixed C, i.e., cardinality of the extracted key-frame set, guarantees summary conciseness to the desired degree.

•The method is more effective in comparison to convex relaxation or iterative approximate solutions to the CSSP, but comes with a greater computational cost.

Evaluation

•Evaluation was performed on a subset of the publicly available IMPART dataset [2], which depicts human subjects performing a series of activities with a static camera, static background and no editing/shot cuts.

•The availability of temporal activity video segmentation ground truth for the IMPART dataset, allows us to perform objective evaluation of the extracted key-frame set, under the following notion: the algorithm, ideally, should extract one key-frame per depicted activity segment.

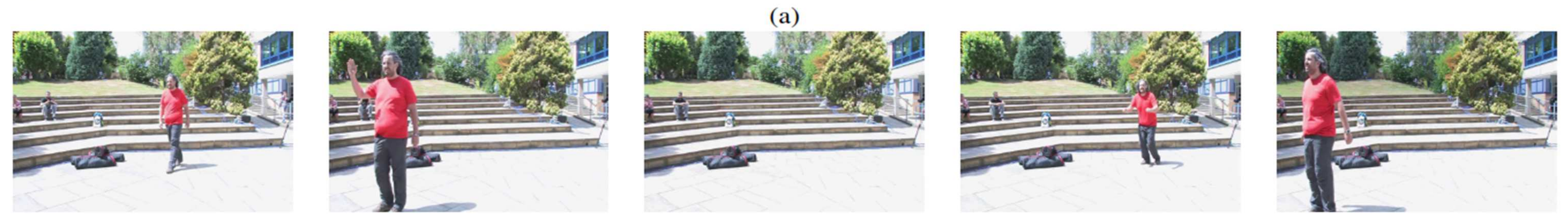
•We propose a relevant, intuitive metric *(Independence Ratio*, IR*)*: the ratio of extracted independent key-frames by the total number of requested key-frames C. Independence of two key-frames implies that they belong to different activity video segments, according to the ground truth.

•Two different video frame representation schemes were tested: Global Histograms and SURF-based visual words. In all cases, four image channels were employed (luminance, color hue, optical flow magnitude, edge map). The related representation vectors were concatenated.

•K-Means++ clustering (a), a typical, straightforward approach to video summarization, was compared to the proposed CSSP-based method (b) in terms of IR performance. The below results were obtained (averaged over the IMPART dataset):

Method	K-Means++	CSSP
Global Histogram	0.571	**0.636**
SURF	0.484	**0.534**

(a)



(b)

Conclusions

•Clustering produces a key-frame set with greater undesired redundancy, while CSSP decomposes the video into disjoint elemental visual word subsets and achieves greater IR score. Low-level global video frame histograms outperform SURF-based BoF representations.

[1] P. Kromer, J. Platos, and V. Snasel, "*Genetic algorithm for the column subset selection problem*", IEEE Complex, Intelligent and Software Intensive Systems (CISIS). 2014, pp. 16–22

[2] T. Theodoridis, A. Tefas and I. Pitas, "*Multi-view semantic temporal video segmentation*", in IEEE International Conference on Image Processing (ICIP). 2016, pp. 3947–3951

Acknowledgement

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).