Unsupervised and Explainable Assessment of Video Similarity

Konstantinos Papoutsakis[1,2#*,] and Antonis A. Argyros[1,2]

[1] Computer Science Department, University of Crete

[2] Institute of Computer Science, FORTH

# Presenting author: Konstantinos Papoutsakis, papoutsa@ics.forth.gr
* Corresponding author: Konstantinos Papoutsakis, papoutsa@ics.forth.gr

ABSTRACT

We propose a novel unsupervised method [1] that assesses the similarity of two videos on the basis of the estimated relatedness of the objects and their behavior, and provides arguments supporting this assessment.

A video is represented as an action graph i.e. a complete, undirected, labeled graph whose nodes correspond to tracked objects. The edges of the graph represent object relations and interactions. The weight of an edge aggregates the dissimilarity of the objects it connects with respect to (i) their semantic similarity (estimated based on the semantic affinity of their labels) and (ii) the dissimilarity of their behavior in time (i.e. estimated based on temporal co-segmentation of their trajectories [2]).

Then, the similarity/distance of a pair of videos is estimated based on an approximation of the Graph Edit Distance (GED) [3] between the corresponding action graphs. GED is a well-known and effective method for inexact graph matching. In the context the proposed methodology it applied to establish meaningful correspondences between the two compared videos, i.e., identifies object pairs that are semantically related, exhibit similar interactions with other objects, or both. Similar actions/interactions are also localized in time based on temporal co-segmentation between the 3D trajectories of two objects [2].

Thus, on-top of estimating a quantitative measure of video similarity, the proposed method establishes spatiotemporal correspondences between objects across videos if these objects are semantically related, if/when they interact similarly, or both, providing explanations of why and when two videos are similar. We consider this an important step towards explainable assessment of video and action similarity.

Our methodology is evaluated using the CAD-120 dataset on the tasks of nearest neighbor action classification and pairwise action matching and ranking, and is shown to compare favorably to state-of-art supervised and unsupervised learning methods.

REFERENCES

[1] Papoutsakis K. and Argyros A. A. 2019. British Machine Vision Conference, Unsupervised and Explainable Assessment of Video Similarity.

[2] Papoutsakis K., Panagiotakis C., and Argyros A. A. 2017. IEEE Computer Vision and Pattern Recognition Conference. Temporal action co-segmentation in 3d motion capture data and videos.

[3] Riesen K. and Bunke H. 2009. Image Vision Computing, 27. Approximate graph edit distance computation by means of bipartite graph matching.