Title: A cognitive based model for event segmentation

Authors: Mariella Dimiccoli, Haoyi Xu, Petia Radeva, University of Barcelona, Spain

Keywords: event segmentation, neural event representation, neural networks

Abstract:

Perceptual information from unconstrained videos, and more in general from the world in which we act, arrives to our brain in a continuous manner over time by our visual system. In spite of that, instead of a continuous visual stream, we do experience this visual information as a sequence of coherent and bounded perceptual units, called *events* in the cognitive literature [1]. Event representation is currently an active area of research in both computer vision and neuroscience. In computer vision, event representation is tightly coupled to event detection and recognition, that are crucial to automatically indexing and retrieving the growing size of today's available videos on internet. In neuroscience, the mechanism underlying event representation and learning are still not well understood. Classically, the surprise at the occurrence of an unpredicted observation has been considered the major cue for event segmentation. However, recent experimental findings [2] have shown that neural representation of events are not tied to predictive uncertainty, but arise from temporal community structures: items that share the temporal context are grouped together in a representational space.

Inspired by these findings, we investigate a new cognitive based model for event segmentation in photo-streams (2fpm) captured by a wearable camera, which exploits this pattern of temporal overlap for dividing an unconstrained stream of images into events. We first detect concepts in each image separately by employing a convolutional neural network approach and later, by leveraging WordNet, we cluster the detected concepts in a semantic space, hence defining a vocabulary of concepts. Each image is therefore represented by a concept vector, whose elements indicate the confidence with which each concept of the vocabulary is detected in the image. Later, by relying on this semantic representation, we train a feed-forward neural network to predict which concept vector would occur in the next image. To simulate the concept vector sequence, we included a number of localist units equal to the number of concept vectors in both the input and output layers. The model modifies connections weights from the current item layer to learn to activate only the possible possible successors for a given concept vector. After training, we use the trained neural network as a fixed predictor and we associate to each image the activation of the hidden layer of its corresponding concept vector after exposure. Finally, we use these activation vectors as new image representation and we apply a hierarchical merging algorithm to find clusters corresponding to events.

Our preliminary results indicate that the newly learned representational space is more suited than a semantic space for event segmentation. The proposed method is similarly accurate as [3] but it involves less parameters and contrarily to [3], which also relies on contextual information, solely rests upon semantic features. Furthermore, while [3] uses a complex clustering method combining the advantages of agglomerative clustering and of a statistical change detector, the proposed method employs a simple hierarchical merging algorithm based on a binary partition tree representation of the photo-streams. Therefore, there is still room for improvement.

References

[1] J.M. Zacks et al. *Event structure in perception and conception*. Psychological bulletin, 127(1): 3-21, 2001.

[2] A. Schapiro et al. *Neural representations of events arise from temporal community structure*. Nature neuroscience 16(4): 486-492, 2013.

[2] M. Dimiccoli et al. *SRclustering: Semantic regularized clustering for egocentric photo stream segmentation.* https://arxiv.org/abs/1512.07143, 2015.