Event representation in egocentric photo-streams: application to temporal segmentation

Mariella Dimiccoli
Computer Vision Center
Edifici O, Campus UAB, 08193 Bellaterra (Cerdanyola del Vallès), Barcelona, Spain
mariella.dimiccoli@cvc.uab.es

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, unpunctuated visual stream, we do experience this visual information as a sequence of coherent and bounded perceptual units having beginnings and ends. In the cognitive literature, these perceptual units have been called *events*. In recent years, event representation has gained a lot of research attention in computer vision. The main motivation behind this trend is the growing size of today's available videos on internet which raises the need of automatically detecting and recognizing the type of complex event occurring in them. However, the detection of complex event in unconstrained videos is an extremely challenging task due to their huge intra-class variation. Most recent and promising approaches for event detection [6, 2] use concept scores as intermediate representation, which is the occurrence confidence for the concepts in the video. For example, the event *Having a dinner with friends* can be described as the occurrence of *food*, *laughing*, *people*, *bottles* etc. However, the resulting concept-based event representation is highly noisy due to the high variability of concept's appearance on the one hand and to the high variability of the concept representation for a complex event on the other hand.

In this work, we explore two different approaches for event representation in a particular kind of unconstrained scenario, the one of egocentric photo-streams captured by a low frame-rate wearable camera, and we evaluate their effectiveness for the task of temporal segmentation. Egocentric photo streams generally appear in form of long unstructured sequences of thousands of images, often with a high degree of redundancy. Due to the very low frame rate of the camera (up to 3fpm) and to the motion of the camera wearer, temporally adjacent images may present abrupt changes in appearance even when the semantic content rests the same. Hence, robust techniques are required to group them into semantically meaningful temporal segments. In the following, we proposes two different approaches for event representation in egocentric photo-streams: one is based on the integration of vision and language and the other one is inspired by recent experimental findings on event representation [5].

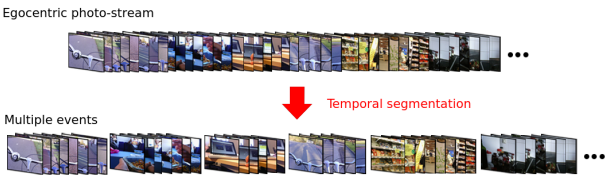

Figure 1. Example of temporal segmentation on an egocentric photo-stream

1. Integrating vision and language for event representation

Focusing on what the camera wearer sees, we consider as an event in egocentric photo-streams a set of temporally adjacent images sharing contextual information represented in terms of semantic visual concepts. However, not all semantic concepts in an image are equally discriminant for what the camera wearer is seeing: concepts like trees and buildings can be more discriminant than concepts like hands or mobile phones, since the former characterizes a specific environment such as forest or street, whereas the latter can be found in many different environments. To take this into account, our feature vector is obtained by concatenating global image features extracted by applying a pre-trained Convolutional Neural Network on the full image with a vocabulary of concepts defined in a semantic space by integrating vision and language processing. More precisely, firstly concepts are detected using state-of-art concept detectors based on convolutional neural networks [4] and, secondly, the concept vectors are clustered in a semantic space taking into account their similarity on WordNet. The centroid of each cluster, defined as the most representative word of the cluster, is considered in the concept vocabulary for event representation. By relying on these seman-

tic features, a graph-cuts optimization strategy is then used to integrate the temporal segmentations produced by ADWIN [1], a change detector method, and the agglomerative clustering, two methods with complementary properties for temporal segmentation. Experiments over egocentric sets of nearly 17,000 images, show that the proposed approach outperforms state-of-the-art methods. More details can be found in [3].

2. A biologically-inspired approach for event representation

Inspired by recent experimental findings on event representation in our brain [5], we trained a three-layer feed-forward neural network which takes as input a concept vector representing the current image to predict which concept vector would occur in the next image. The model modified connections weights from the current item layer to learn to activate only the possible possible successors for a given concept vector. We then applied a very simple unsupervised clustering method, mainly k-means, on the hidden representation and we achieved better results for temporal segmentation with respect to the results obtained with the method described in section 1. These results indicate that instead of relying on a concept vector's representation and using a complex model to cluster images into unitary percepts, it is possible to learn more sophisticated event representations and then performing clustering in a simpler and effective way.

These new methods for event representation could be readily fed into the frameworks event recognition in unconstrained videos to further improve their performance.

References

[1] R. G. A. Bifet. Learning from time-changing data with adaptive windowing. In *In Proc. SIAM International Conference on Data Mining*, 2007.

[2] X. Chang, Y. Yang, A. G. Hauptmann, E. P. Xing, and Y.-L. Yu. Semantic concept discovery for large-scale zero-shot event detection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[3] M. Dimiccoli, M. Bolaños, E. Talavera, M. Aghaei, S. G. Nikolov, and P. Radeva. Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation. *arXiv preprint arXiv:1512.07143*, 2015.

[4] J. Hoffman, S. Guadarrama, E. S. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. Lsda: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*, pages 3536–3544, 2014.

[5] A. C. Schapiro, T. T. Rogers, N. I. Cordova, N. B. Turk-Browne, and M. M. Botvinick. Neural representations of events arise from temporal community structure. *Nature Neuroscience*, 16(4):486–492, 2013.

[6] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative cnn video representation for event detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1798–1807, 2015.