Exemplar-based Action Recognition in Video

Geert Willems[1]
homes.esat.kuleuven.be/~gwillems
Jan Hendrik Becker[1]
homes.esat.kuleuven.be/~jhbecker
Tinne Tuytelaars[1]
homes.esat.kuleuven.be/~tuytelaa
Luc Van Gool[12]

[1] ESAT-PSI/Visics
K.U. Leuven
[2] Computer Vision Laboratory
BIWI/ETH Zürich

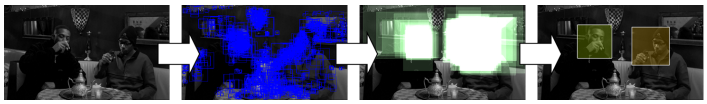

Figure 1: An illustration of the detection pipeline for the 'drinking' action: (1) a frame from the input video, (2) spatio-temporal features are computed throughout the video, (3) exemplar-based hypotheses are generated and scored, (4) a grouping algorithm yields two distinct detections.

Over recent years, a lot of progress has been made towards automatic annotation of video material, especially in the context of object and scene recognition. However, in comparison, action recognition is still in its infancy. Whereas originally silhouette-based approaches or approaches based on pose estimation have been studied mostly, good results have been reported recently using extensions of traditional object recognition approaches to the spatio-temporal domain [2, 3, 5]. These methods consider actions as typical spatio-temporal patterns that can be modeled using local features, optical flow, or gradient-based descriptors. It is in this line of research that our work is situated. More specifically, we build on the work of Chum *et al*. [1] which is an exemplar-based approach for object detection using local features that can be situated somewhere in between sliding window based approaches and the *Implicit Shape Model* (ISM) [4].

We extend the exemplar-based object detection work of Chum *et al*. [1] to the spatio-temporal domain where we use the recently proposed local, dense, scale-invariant spatio-temporal features [6]. The overall pipeline is shown in figure 2.

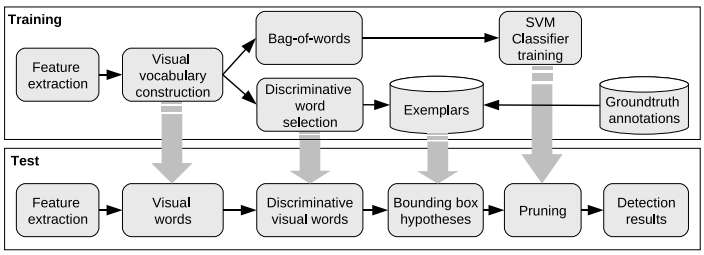

Figure 2: Overview of the exemplar-based detection pipeline.

To the best of our knowledge, we are the first to extend the exemplar-based approach using local features into the spatio-temporal domain. This allows us to avoid the problems that typically plague sliding window based approaches – in particular the exhaustive search over spatial coordinates, time, and spatial as well as temporal scales. Next, we briefly describe both the training and detection pipeline.

Training For the training, we start with a set of annotated videos and extract local spatio-temporal features. After clustering, we select the top N most discriminative visual words based on a F_β-measure. Exemplars are collected for each of the N words, by searching for all features belonging to that word and storing the coordinates of the annotation region relative to each feature's position and scale. Unlike Chum *et al*. [1], we do not cluster the bounding boxes, but keep all of them instead (similar to what is done in the ISM model). At the same time, we also learn a classifier, based on the ground truth annotations in the training data as well as a set of randomly generated negatives (not overlapping with any of the ground truth bounding boxes). We compute a bag-of-words for each annotation and train a non-linear support vector machine using the χ^2-kernel.

Detection The detection pipeline extracts all local spatio-temporal features within the video, keeps those belonging to the top N most discriminative visual words, and generates a set of hypotheses. A first pruning step removes hypotheses that are not deemed useful because of their bounding box properties. We compute the bag-of-words from the features inside each hypothesis' bounding box and assign them a confidence based on the decision value of the previously trained SVM classifier. A second pruning step further removes all hypotheses with a confidence value below a pre-defined threshold. Finally, a simple greedy algorithm is used to merge all hypotheses into several detections. As movie segments, like the test video, typically contain many shot cuts, we do not allow hypotheses to merge together across shot cuts. The different steps in the pipeline are illustrated in figure 1.

We evaluate our approach on the *DrinkingSmoking* dataset from [3] and a novel *ExtendedDrinkingSmoking* dataset, which includes the annotations of 3 additional movies. The latter has also been made publicly available. On the *DrinkingSmoking* dataset, we achieve an Average Precision of 45.2% for the 'Drinking' action, outperforming the best result from [3] by 3%. The top 15 of those detections are shown in figure 3.

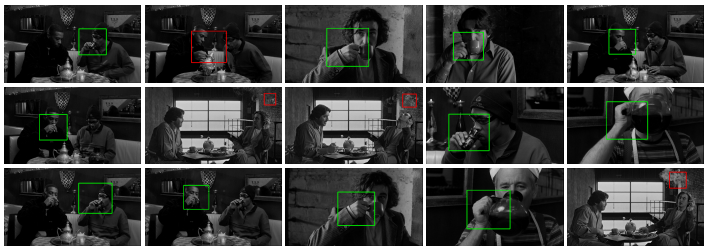

Figure 3: The top 15 detection results for 'Drinking' action on the *DrinkingSmoking* dataset (top to bottom, left to right). True positives are shown in green, false positives in red.

Details of the implemented pipeline together with a quantitative evaluation are described in the paper. We report state-of-the-art results on challenging datasets, extracted from real movies, for both classification and localization. Although the approach has been stripped of any refinements that may boost performance further, the results clearly demonstrate its viability.

[1] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[2] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *International workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.

[3] I. Laptev and P. Perez. Retrieving actions in movies. In *Proceedings ICCV07*, 2007.

[4] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77(1-3):259–289, 2008.

[5] K. Mikolajczyk and H. Uemura. Action recognition with motion-appearance vocabulary forest. In *Proceedings CVPR08*, 2008.

[6] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings ECCV08*, 2008.