Randomized Object-Centric Spatio-Temporal Pyramids for Egocentric Activity Recognition

BMVC 2013 Submission # 131

Egocentric video and wearable computing have become increasingly prevalent in the past decade, resulting in a huge explosion in the amount of available video content and increased attention from the computer vision community. However, existing methods for activity recognition often use predefined spatio-temporal binning schemes to aggregate features. This encodes information beyond what is possible with a pure "bag of words" model, but is ultimately inflexible and may fail to capture important spatio-temporal relationships between features. We propose to randomly generate a pool of candidate binning schemes and use a boosting algorithm to combine those which are most discriminative. In order to efficiently focus the candidate partition schemes, we create biased partitions using "object-centric" cuts in video volumes. Partition schemes generated using our method have a high probability of cutting through video regions that contain "active objects," the objects being interacted with by the user during a given frame. Given a set of training videos, our method first computes histograms of active object locations across each (x,y,t) dimension, then uses these histograms to generate a pool of object-centric partition schemes that have a high probability of cutting through regions that often contain active objects. We use a boosting algorithm to learn which partitioning schemes are most discriminative and form a final strong classifier. Our main novel contribution is two-fold: we show how to learn the most useful partition schemes in an egocentric setting, and we focus candidate partition schemes by exploiting locations of active objects. Our approach yields state-of-the-art recognition performance, and we find that object-centric partition schemes are often more discriminative than their unbiased counterparts.

Egocentric activities are well-defined by the types of objects that are interacted with by users during particular actions ("active objects") [4], yet how to optimally aggregate features across space-time remains unclear. The familiar bag-of-words approach can be used to aggregate features with reasonable performance, but ultimately falls short because it fails to capture temporal dependencies between features. The pyramid is a well-known extension of a pure bag-of-words model that encodes spatial relationships between features by recursively subdividing images or video and extracting features from each spatial bin [3], yielding impressive results across a range of applications. Existing methods for activity recognition often rely on hand-coded partition schemes [1, 2, 4]. With a small pool of hand-coded schemes for imposing spatial information, the most discriminative space-time relationships between features may not be

Our idea is to randomly generate a pool of candidate partitioning schemes. We then aggregate spatio-temporal features in a learned way, using a boosting algorithm to select those partitioning schemes which are most discriminative. captured.

Our algorithm takes as input a collection of N labeled training videos where (V_i, c_i) denotes a video clip and its associated ground-truth activity label, and a pool of M candidate partition patterns $\{\theta_1, \theta_2, ..., \theta_M\}$. We use the output of the aforementioned object detectors trained on composite object models as our features to be pooled.

We evaluate the performance of our method using a cross-validation experiment and find that our method using object-centric pyramids improves upon the current state of the art. Furthermore, we show that an increase in discriminative over unbiased pools is most visible at small pool sizes. This suggests that it is possible to use a small pool of object-centric partitions instead of a large pool of unbiased partitions and still obtain good classification results.

Our main novel contribution is two-fold. We show how to learn the most discriminative partition schemes for spatio-temporal binning in video feature space, and we introduce object-centric partition schemes, which have a high probability of cutting through video regions known to frequently contain active objects. Unlike previous work, we randomly generate a pool of candidate partitioning schemes and select those which are most discriminative using a boosting algorithm. Our recognition approach improves on the current state of the art, and our experiments demonstrate the positive impact of taking active object locations into account by generating object-centric partition schemes.

[1] Jaesik Choi, Won J. Jeon, and Sang-Chul Lee. Spatio-temporal pyramid matching for sports videos. *MIR*, 2008.

[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. *CVPR*, 2008.

[3] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *CVPR*, 2006.

[4] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. *CVPR*, 2012.