Object-Centric Spatio-Temporal Pyramids for Egocentric Activity Recognition

Tomas McCandless
tomas@cs.utexas.edu
Kristen Grauman
grauman@cs.utexas.edu

Department of Computer Science,
University of Texas at Austin

Motivation Egocentric computer vision entails analyzing images and video that originate from a wearable camera, which is typically mounted on the head or chest. Seeing the world from this first-person point of view affords a variety of exciting new applications and challenges, particularly as today's devices become increasingly lightweight and power efficient. Nearly all applications of egocentric vision demand robust methods to recognize activities as seen from the camera wearer's perspective.

High-level representations based on detected objects are a promising way to encode video clips when learning egocentric activities [1, 3, 5]. In particular, recent work explores a "bag-of-objects" histogram of all objects detected in a video sequence, as well as an extension that captures the objects' relative temporal ordering [5]. Similarly, for third-person video, activity recognition methods often employ histogram pyramids with regular grid structures to pool local descriptors (e.g., [2, 4]). Coupled with standard discriminative classifiers, this representation shows very good recognition results.

Problem However, existing methods that pool localized visual features into space-time histogram bins do so using hand-crafted binning schemes. The problem with defining the spatio-temporal bins *a priori* is that they may not offer the most discriminative representation for the activity classes of interest. That is, the hand-crafted histogram bins may fail to capture those space-time relationships between the local features that are most informative for activity recognition.

Our idea To overcome this limitation, we propose to *learn* discriminative spatio-temporal histogram partitions for egocentric activities. Rather than manually define the bin structure, we devise a boosting approach (Figure 1) that automatically selects a small set of useful spatio-temporal pyramid histograms among a randomized pool of candidate partitions. To minimize a computationally expensive search, we further propose a way to meaningfully bias the partitions that comprise the candidate pool. We devise an *object-centric* cutting scheme that prefers sampling bin boundaries near objects involved in the egocentric activities. As a result, we focus the randomized pool of space-time partitions to the egocentric setting while also improving training efficiency.

Approach Our approach works as follows. Given a set of egocentric training videos labeled according to their activity class, we first run object detectors on the frames to localize any objects of interest—both those that are "passive" and those that are "active" in an interaction with the camera wearer. Active objects are those being manipulated by the user [5], while passive objects are those which lie inert in the frame background. We then construct a series of candidate space-time pyramids. Given this candidate pool, we compute the corresponding series of object histograms for each training video. Then, we apply multi-class boosting to select a subset of discriminative pyramid structures. At the end, we have a strong classifier that can predict the activity labels of new videos, using only those randomized pyramids selected by the learning algorithm.

While boosting gives an automated way to select informative pyramids, its training time depends linearly on the number of candidates we include in the pool. To avoid an excessive search, we focus the candidate pool in a way that is meaningful for egocentric data. Rather than sample cuts uniformly at random, our idea is to sample the cuts according to the distribution of active objects as they appear in the training videos (Figure 2). We refer to these as *object-centric cuts* (OCC).

Result summary We validate our method on the Activities of Daily Living (ADL) dataset [5], a large and realistic dataset consisting of 18 household actions, like doing laundry, washing dishes, etc. We compare our boosted RSTP+OCC approach to four baselines: 1) a standard bag-of-words (BoW) on space-time interest points with HoG/HoF visual words, 2) a bag-of-objects approach that replaces visual words with detected objects, 3) the Temporal Pyramid (TempPyr) developed in [5], which represents the state of the art performance for this dataset, and 4) a boosting approach just like the proposed method, but without the object-centric cuts (denoted RSTP, for randomized space-time pyramids).

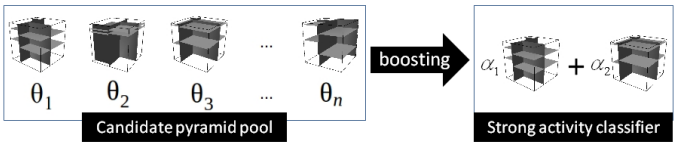

Figure 1: We take a pool of randomized space-time pyramids with object-centric cuts, and use boosting to select those that are most discriminative for egocentric activity recognition. Each pyramid is used to histogram the object occurrences across space and time, revealing the primary interactions of the camera wearer.

(a) Object-centric cuts

(b) Uniformly random shifts

Figure 2: Example partitions using either object-centric (a) or uniformly sampled randomized cuts (b). For display purposes we show cuts on 2D frames, but all cuts are 3D in space-time. The proposed object-centric cuts focus histograms on human-object interactions.

As shown in Table 1, our approach outperforms all four baselines and improves the state of the art. Notably, our proposed object-centric cuts are essential for our strong recognition result. Simply using boosting with purely randomized partitions (RSTP) is noticeably weaker. This shows it is useful to bias bins according to object interactions for egocentric data.

Main contribution Our main contribution is two-fold. We show how to learn the most discriminative partition schemes for spatio-temporal binning in action recognition, and we introduce object-centric cuts for egocentric data. Our proposed approach improves on the current state of the art for recognizing activities of daily living from the first person viewpoint.experiments demonstrate the positive impact of taking active object locations into account via object-centric cuts.

BoW	Bag-of-objects	TempPyr [5]	RSTP	RSTP+OCC
16.5%	34.9%	36.9%	33.7%	**38.7%**

Table 1: Overall classification accuracy on ADL. Our method, RSTP+OCC, improves the state of the art.

[1] A. Behera, D. C. Hogg, and A. G. Cohn. Egocentric activity monitoring and recovery. In *ACCV*, 2012.

[2] J. Choi, W. Jeon, and S.-C. Lee. Spatio-temporal pyramid matching for sports videos. In *ACM Multimedia*, 2008.

[3] A. Fathi, A. Farhadi, and J. Rehg. Understanding egocentric activities. In *ICCV*, 2011.

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

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