Recognizing activities with cluster-trees of tracklets

Adrien Gaidon
http://lear.inrialpes.fr/people/gaidon
Zaid Harchaoui
http://lear.inrialpes.fr/people/harchaoui
Cordelia Schmid
http://lear.inrialpes.fr/people/schmid

LEAR - INRIA Grenoble, LJK
655, avenue de l'Europe
38330 Montbonnot, France

Although the structure of simple actions can be captured by rigid grids [3] or by sequences of short temporal parts [2], *activities* are composed of a variable number of sub-events connected by more complex spatio-temporal relations. In this paper, we learn how to automatically represent activities as a hierarchy of mid-level motion components in order to improve activity classification in real-world videos. This hierarchy is a video-specific, data-driven decomposition obtained by clustering *tracklets*, *i.e.* local point trajectories of a fixed small duration.

Our first contribution is a hierarchical spectral clustering algorithm, based on top-down recursive bi-partitioning. We propose to robustly threshold tracklet projections on an approximate spectral embedding by minimizing a spatio-temporal *connectivity* cost. This allows for clusters of arbitrary shape and an efficient greedy splitting strategy that automatically determines the number of clusters. The resulting hierarchical decomposition provides structural information relating motion parts together.

Our second contribution is the use of this entire tree structure, called *cluster-tree* (*c.f.* Figure 1), in order to build a hierarchical model of the motion content of a video. We introduce a corresponding tree representation of actions, called *BOF-tree*. The BOF-tree of a video has the same structure as its cluster-tree and each node is modeled by a bag-of-features (BOF) over the MBH descriptors [6] of its constitutive tracklets. Efficiently using this structural information is challenging as cluster-trees have a variable number of nodes and a structure specific to each video. Furthermore, there is no natural left-to-right ordering of the two children of a parent node. Therefore, we introduce an efficient positive definite kernel — called the "All Tree Edge Pairs" (ATEP) kernel — that computes the structural and visual similarity of two hierarchical decompositions by relying on models of their parent-child relations as described in the following. Let $\mathcal{T}_i = (\mathcal{V}_i, \mathcal{E}_i)$, $i \in \{1,2\}$, be two BOF-trees, defined from their set of vertices (nodes) $\mathcal{V}_i$ and directed edges (parent-child relations) $\mathcal{E}_i$. Each node $v \in \mathcal{V}_i$ is represented by a BOF — noted $b[v]$ — over its constitutive tracklets. We model a directed edge $e = (v_p, v_c) \in \mathcal{E}_i$ by the concatenation — noted $b[e] = (b[v_p], b[v_c])$ — of the BOF of the child node v_c with the BOF of its parent node v_p. Let h be a kernel between BOF, $r_i \in \mathcal{V}_i$ be the root of $\mathcal{T}_i$, and $w_r \in (0,1)$ a cross-validated parameter encoding a prior on the importance of the root-to-root comparisons. Our ATEP kernel is defined as:

$$k(\mathcal{T}_1, \mathcal{T}_2) = w_r \cdot h(b[r_1], b[r_2]) + \frac{1-w_r}{|\mathcal{E}_1||\mathcal{E}_2|} \cdot \sum_{\substack{e_1 \in \mathcal{E}_1 \\ e_2 \in \mathcal{E}_2}} h(b[e_1], b[e_2])$$

As described in the paper, this kernel can be seen as a weighted similarity between all sub-trees of two BOF-trees.

We use our ATEP kernel in conjunction with SVM classifiers on videos represented by BOF-trees, *i.e.* hierarchically structured sets of motion components. We present experimental results on two recent challenging benchmarks focusing on complex activities: the Olympics Sports dataset [4] and the human-human interactions of the High Five dataset [5]. Table 1 reports performance comparisons between baselines (described in the legend), the state of the art and our method. Our hierarchical ATEP kernel on SDT BOF-trees improves over the unstructured BOF baselines. This confirms the importance of leveraging structure information to recognize complex activities. Note, however, that only our method yields clear performance improvements, whereas other structured baselines are less accurate. This shows that properly decomposing activities is a challenging problem that is critical for performance. In particular, using hierarchical relations between motion components with our ATEP kernel consistently improves over the "flat" baselines relying on unrelated sets of clusters such as the leaves of cluster-trees or clusters obtained by k-means. Table 1 also shows that the BOF-trees produced by our SDT algorithm yield more powerful models than the SDKM ones obtained by bi-partitioning k-means. Finally, our approach outperforms the state of the art on both datasets, including latent part models [4] (+10.6%), complex graphical models resulting from video segmentation [1] (+5.4%), and interaction-specific structured learning [5] (+22.8%).

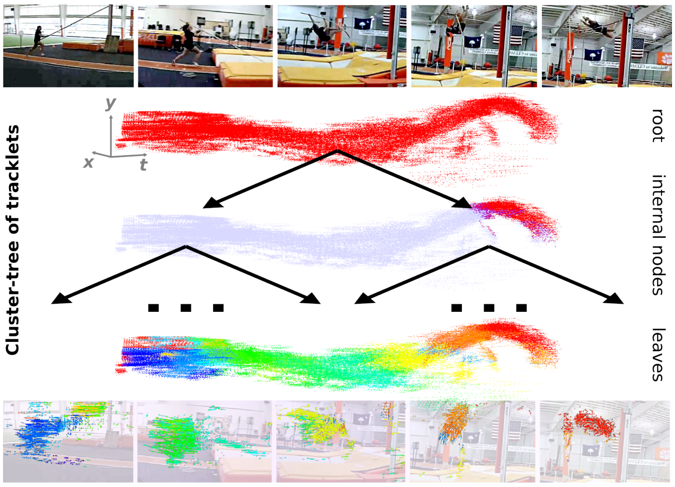

Figure 1: Example of a hierarchical motion decomposition obtained by our divisive clustering algorithm on dense tracklets. Our approach uses the whole cluster-tree to compare videos.

SDT trees with ATEP	**82.7**
SDKM trees with ATEP	76.7
SDT leaves	77.9
SDKM leaves	72.2
spectral	71.7
kmeans	70.8
Wang *et al.* [6]	75.9
Laptev *et al.* [3]	61.3
Brendel and Todorovic [1]	77.3
Niebles *et al.* [4]	72.1

(a) Olympics Sports (Accuracy in %)

SDT trees with ATEP	**55.6**
SDKM trees with ATEP	53.8
SDT leaves	49.5
SDKM leaves	48.6
spectral	48.9
kmeans	50.1
Wang *et al.* [6]	53.4
Laptev *et al.* [3]	36.9
Patron-Perez *et al.* [5]	32.8

(b) High Five (AP in %)

Table 1: Performance on the Olympics Sports [4] and High Five [5] datasets. Results with our Spectral Divisive Thresholding algorithm are noted "SDT". We compare with the state of the art, BOF baselines [3, 6], "flat" decompositions obtained by k-means and spectral clustering, as well as hierarchical decompositions obtained by a baseline spectral divisive bi-partitioning k-means algorithm noted "SDKM".

Acknowledgments This work was partially funded by the MSR/INRIA joint project, the European project AXES and the PASCAL 2 Network of Excellence.

[1] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *ICCV*, 2011.

[2] A. Gaidon, Z. Harchaoui, and C. Schmid. Actom sequence models for efficient action detection. In *CVPR*, 2011.

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

[4] J.C. Niebles, CW. Chen, , and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010.

[5] A. Patron-Perez, M. Marszałek, A. Zisserman, and I. D. Reid. High five: Recognising human interactions in TV shows. In *BMVC*, 2010.

[6] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. In *CVPR*, 2011.