Unsupervised Learning and Segmentation of Complex Activities from Video

Fadime Sener, Angela Yao
University of Bonn, Germany
{sener,yao}@cs.uni-bonn.de

This paper presents a new method for unsupervised segmentation of complex activities from video into multiple steps, or sub-activities, without any textual input [7]. A complex activity is a procedural task with multiple steps or sub-activities that follow some loose ordering. Complex activities can be found in instructional videos; YouTube hosts hundreds of thousands of such videos on activities as common as *'making coffee'* to the more obscure *'weaving banana fibre cloths'*. We propose an iterative discriminative-generative approach which alternates between discriminatively learning the appearance of sub-activities from the videos' visual features to sub-activity labels and generatively modelling the temporal structure of sub-activities using a Generalized Mallows Model [2]. In addition, we introduce a model for background to account for frames unrelated to the actual activities. Our approach is validated on the challenging Breakfast Actions [4] and Inria Instructional Videos [1] datasets and outperforms both unsupervised and weakly-supervised state of the art.

The iterative model we propose alternates between learning a discriminative representation of a video's visual features to sub-activities and a generative model of the sub-activities' temporal structure. By combining the sub-activity representations with the temporal model, we arrive at a segmentation of the video sequence, which is then used to update the visual representations. We represent sub-activities by learning linear mappings from visual features to a low dimensional embedding space with a ranking loss. The mappings are optimized such that visual features from the same sub-activity are pushed together, while different sub-activities are pulled apart.

Temporally, we treat a complex activity as a sequence of permutable sub-activities and model the distribution over permutations with a Generalized Mallows Model (GMM) [2]. In our method, the GMM assumes that a canonical sequence ordering is shared among videos of the same complex activity. There are several advantages of using the GMM for modelling temporal structure. First and foremost, the canonical ordering enforces a global ordering constraint over the activity – something not possible with Markovian models [4, 5, 8] and recurrent neural networks (RNNs) [11]. Secondly, considering temporal structure as a permutation offers flexibility and richness in modelling. We can allow for missing steps and deviations, all of which are characteristic of complex activities, but cannot be accounted for with works which enforce a strict ordering [1]. Finally, the GMM is compact – parameters grow linearly with the number of sub-activities, versus quadratic growth in pairwise relationships, *e.g.* in HMMs.

Within a video, it is unlikely that every frame corresponds to a specified sub-activity; they may be interspersed with unrelated segments of actors talking or highlighting previous or subsequent sub-activities. Depending on how the video is made, such segments can occur arbitrarily. In this paper we extend our segmentation method to explicitly learn about and represent such "background frames" so that we can exclude them from the temporal model.

Proposed Work: We are the first to explore a fully unsupervised method for temporal understanding of complex activities in video without requiring any text. We design a discriminative appearance learning model to enable the use of GMMs on state-of-the-art visual features [6, 9, 10].

Assume we are given a collection of M videos, all of the same complex activity, and that each video is composed of an ordered sequence of multiple sub-activities. A single video i with J_i frames can be represented by a design matrix of features $\mathbf{F}_i \in \mathbb{R}^{J_i \times D}$, where D is the feature dimension.

We first describe how we discriminatively learn the features $\mathbf{F}$. Within a video collection of a complex activity there may be huge variations in visual appearance, even with state-of-the-art visual feature descriptors. Suppose for frame j of video i we have video features $\mathbf{X}_{ij}$. These features, if clustered naively, are most likely to group together according to video rather than sub-activity. To cluster the features more discriminantly, we learn a linear mapping of these features into a latent embedding space. We also define in the latent space K anchor points, with locations determined by a second mapping. Our objective in learning the embeddings is to cluster the video features discriminatively. We achieve this by encouraging the $\mathbf{X}_{ij}$ belonging to the same sub-activity to cluster closely around a single anchor point while being far away from the other anchor points. If

we assign each anchor point to a given sub-activity, then we can learn the embedding parameters by minimizing a pairwise ranking loss. The loss encourages the distance of $\mathbf{X}_{ij}$ in the latent space to be closer to the anchor point k^* associated with the true sub-activity than any other anchor point by a margin.

After discriminatively learning the features $\mathbf{F}$ we describe our standard temporal model. Given a collection of M videos of the same complex activity, we would like to infer the sub-activity assignments $\mathbf{z} = \{\mathbf{z}_i\}, i \in \{1, \ldots, M\}$. For video i, $\mathbf{z}_i = \{z_{ij}\}, j \in \{1, \ldots, J_i\}$, $z_{ij} \in \{1, \ldots, K\}$ can be assigned to one of K possible sub-activities. We introduce $\mathbf{a}_i$, a bag of sub-activity labels for video i, *i.e.* the collection of elements in $\mathbf{z}_i$ but without consideration for the temporal frame ordering. The ordering is then described by $\boldsymbol{\pi}_i$. $\mathbf{a}_i$ is expressed as a vector of counts of the K possible sub-activities, while $\boldsymbol{\pi}_i$ is expressed as an ordered list. We model $\mathbf{a}$ as a multinomial, with parameter $\boldsymbol{\theta}$ and a Dirichlet prior with hyperparameter θ_0. For the ordering $\boldsymbol{\pi}$, we use a GMM with the exponential prior and hyperparameters ρ_0 and ν_0. Our interest is to infer the posterior $P(\mathbf{z}, \boldsymbol{\rho} | \mathbf{F}, \theta_0, \rho_0, \nu_0)$ for the entire video corpus. Directly working with this posterior is intractable, so we make MCMC sampling-based approximations. Specifically, we use slice sampling for $\boldsymbol{\rho}$ and collapsed Gibbs sampling [3] for $\mathbf{z}$. Since $\mathbf{z}$ is fully specified by $\mathbf{a}$ and $\boldsymbol{\pi}$, it is equivalent to sample $\mathbf{a}$ and $\boldsymbol{\pi}$.

To consider background, we extend the label assignment vector $\mathbf{z}$ with a binary indicator variable $b_{ij} \in \{0, 1\}$ for each frame. The indicator b_{ij} follows a Bernoulli variable parameterized by λ, with a beta prior, *i.e.* $\lambda \sim \text{Beta}(\alpha, \beta)$. In this setting, $\mathbf{z}_i$ is determined by the bag of sub-activities $\mathbf{a}_i$, the ordering $\boldsymbol{\pi}_i$, and background vector $\mathbf{b}_i = \{b_{ij}\}$, where $\mathbf{b}_i$ indicates the frames to be excluded from sub-activity consideration. For example, for video i, given $\mathbf{a}_i = [6\ 3\ 5]$, $\boldsymbol{\pi}_i = [2\ 3\ 1]$ and $\mathbf{b}_i = [11100111001100011110011]$, the sub-activity assignment is $\mathbf{z}_i = [22200333003300011110011]$.

Experiments: We demonstrate that our method achieves competitive results comparable to or better than the state of the art on two challenging complex activity datasets, Breakfast Actions [4] and Inria Instructional Videos [1]. The Breakfast Actions has 1,712 videos and no background frames. We find that our fully unsupervised approach has performance that is state of the art. Inria Instructional Videos contains 150 narrated videos of 5 complex activities collected from YouTube. The videos are labelled, including the background. Our performance across the five activities is consistent and varies much less than [1]. In general, we attribute our stronger performance to the fact that the GMM can model flexible sub-activity orderings, while [1] enforces a strict ordering.

Conclusion: In this paper we present an unsupervised method for partitioning complex activity videos into coherent segments of sub-activities. We learn a function assigning sub-activity scores to a video frame's visual features, we model the distribution over sub-activity permutations by a Generalized Mallows Model (GMM). Furthermore, we account for background frames not contributing to the actual activity.

We successfully test our method on two datasets of this challenging problem and are either comparable to or outperform the state of the art, even though our method is completely unsupervised, in contrast to the existing work. Our method is able to produce coherent segments, at the same time being flexible enough to allow missing steps and variations in ordering. Performance drops slightly for complex activities including repetitive sub-activities, as the GMM does not allow for such repeating structures.

Acknowledgments Research in this paper was supported by the DFG project YA 447/2-1 (DFG Research Unit FOR 2535 Anticipating Human Behavior).

References

[1] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016.

[2] M. A. Fligner and J. S. Verducci. Distance based ranking models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 359–369, 1986.

[3] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl 1):5228–5235, 2004.

[4] H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, 2014.

[5] A. Richard, H. Kuehne, and J. Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *CVPR*, 2017.

[6] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013.

[7] F. Sener and A. Yao. Unsupervised learning and segmentation of complex activities from video. *arXiv preprint arXiv:1803.09490*, 2018.

[8] O. Sener, A. R. Zamir, S. Savarese, and A. Saxena. Unsupervised semantic parsing of video collections. In *ICCV*, 2015.

[9] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[10] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103(1):60–79, 2013.

[11] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *CVPR*, 2016.