Action Recognition by Hierarchical Sequence Summarization

Yufeng Jiang

1. Related Work

Learning from a hierarchical feature representation has been a recurring theme in action recognition [6, 12, 3]. One approach uses the popular bag-of-words approach, which detects spatio-temporal interest points (STIP) [4]at local video volumes, constructs a bag-of-words represen- tation of HOG/HOF features extracted around STIPs, and learns an SVM classifier to categorize actions [5]. This has been used to construct a hierarchical feature representation that is more discriminative and context-rich [12, 3]. Sun *et al.* [12] defined three levels of context hierarchy with SIFT-based trajectories, while Wang *et al.* [13] learned interactions within local contexts at multiple spatio-temporal scales. Kovashaka and Grauman [3] proposed to learn class conditional visual words by grouping local features of motion and appearance at multiple space-time scales. While these approaches showed significant improvements over the local feature representation, they use non-temporal machine learning algorithms to classify actions, limiting their application to real-world scenarios that exhibit complex temporal structures.

2. Complexity Analysis

To see the effectiveness of the gate functions, consider another definition of the observation feature function, one without the gate functions,

$$f^1(\mathbf{h}, \mathbf{x}, t; \mathbf{w}) = \frac{1}{|c(\mathbf{x}_t)|} \sum_{\mathbf{x}'} \sum_{d} w_{h,d} \mathbb{1}[h_t = h'] x'_d \quad (1)$$

This does not have the automatic feature learning step, and simply represents the feature as an average of the linearcombinations of features x'_d and weights $w_{h,d}$. As evidenced by the deep learning literature [1], the step of nonlinear feature learning leads to a more discriminative representation.

Our model parameter vector is $w = [w_{g,h}; w_{g,d}; w_{y,h}; w_{y,h,h}]$ and has the dimension of $GH + GD + Y\ H + Y\ HH$, with the number of gate functions G, the number of latent states H, the feature dimension D, and the number of class labels Y. Given a chain-structured sequence x of length T, we can solve the inference problem at $O(YTH^2)$ using a belief propagation algorithm.

2.1. Sequence Summarization

There are many ways to summarize $\mathbf{x}^l$ to obtain a temporally coarser-grained sequence summary $\mathbf{x}^{l+1}$. One simple approach is to group observations from $\mathbf{x}^l$ at a fixed time interval, *e.g.*, collapse every two consecutive observations and obtain a sequence with half the length of $\mathbf{x}^l$. However, as we show in our experiments, this approach may fail to preserve important local information and result in over-grouping and over-smoothing.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ be a weighed graph at the l-th layer, where $\mathcal{V}$ is a set of nodes (latent variables), $\mathcal{E}$ is a set of edges induced by a linear chain, and $\mathcal{W}$ is a set of edge weights defined as the similarity between two nodes. The algorithm produces a set of super observations $\mathcal{C} = \{c(\mathbf{x}_1^{l+1}, \ldots, c(\mathbf{x}_T^{l+1}))\}$.

The algorithm merges $c(\mathbf{x}_s^{l+1}$ and $c(\mathbf{x}_t^{l+1})$ if the difference between the groups is smaller than the minimum internal difference within the groups. Let teh internal difference of a group c be $Int(c) = max_{(s,t) \in mst(c, \mathcal{E}_c)} w_{st}$, *i.e.*, the largest weight in the minimum spanning tree of the group c with the corresponding edge set $\mathcal{E}_c$. The minimum internal difference between two groups c_s and c_t is defined as $MInt(c_s, c_t) = min(Int(c_s) + \tau(c_s), Int(c_t) + \tau(c_t))$ where $\tau(c_s) = \tau/|c_s|$ is a threshold function; it controls the degree to which the difference two groups must be greater than their internal differences in order for there to be evidence of a boundary between them.

Similarity Metric: They define the similarity between two nodes (*i.e.*, the weight w_{st}) as [9]

$$w_{st} = \sum_{y,h'} |p(h_s = h'|y, \mathbf{x}; \mathbf{w}) - p(h_t = h'|y, \mathbf{x}; \mathbf{w}| \quad (2)$$

that is , it is the sum of absolute differences of the posterior probabilities between the two corresponding latent variables, marginalized over the class label.[1]

[1] Other metrics can also be defined in the latent space. We experimented with different weight functions, but the performance difference was not significant. We chose this definition because it performed well across different datasets and is computationally simple.

Model	Mean Accuracy
HMM (from [8])	84.83%
CRF (from [8])	86.03%
MM-HCRF (from [10])	93.79%
Quattoni *et al.* [11]	93.81%
Shyr *et al.* [7]	95.30%
Song *et al.* [10]	97.79%
HCNF	97.79%
Our HSS Model	**99.59%**

Table 1. Experimental results from the ArmGesture dataset.

Model	Mean Accuracy
SVM (from [2])	51.89%
HMM (from [2])	52.29%
Bousmails *et al.* [2]	64.22%
Song *et al.* [11]	71.99%
HCNF	73.35%
Our HSS Model	**75.56%**

Table 2. Experimental results from the Canal9 dataset.

2.2. The HSS Model

They formulate their model, Hierarchical Sequence Summarization (HSS), as the conditional probability distribution

$$p(y|\mathbf{x};\mathbf{w}) \propto p(y|\mathbf{x}^1,\ldots,\mathbf{x}^{\mathcal{L}};\mathbf{w}) \propto \prod_{l=1}^{\mathcal{L}} p(y|\mathbf{x}^l;\mathbf{w}^1). \tag{3}$$

Note the layer-specific model parameter vector $\mathbf{w}^l$, $\mathbf{w} = [\mathbf{w}^1,\ldots,\mathbf{w}^{\mathcal{L}}]$.

The first derivation comes from their reformulation of $p(y|\mathbf{x};\mathbf{w})$ using hierarchical seqence summaries, the second comes from the way they construct the sequence summaries. To see this, recall that they obtain a sequence summary $\mathbf{x}^{l+1}$ given the posterior of latent variables $p(\mathbf{h}^l|y,\mathbf{x}^l;\mathbf{w}^l)$, and the posterior is computed based on the parameter vector $\mathbf{w}^l$. To make their model tractable, they assume that a parameter vector at each layer $\mathbf{w}^l$ is independent of each other. As a result, they can express the second term as the product of $p(y|\mathbf{x}^l;\mathbf{w}^l)$.

2.3. Results

Table 1 and Table 2used at [9]shows experimental results on the ArmGesture and Canal9 datasets, respectively. They include previous results on each dataset reported in the literature; They also include the result obtained by their using CNF [7] with latent variables (HCNF). As can be seen, their approach outperforms all the state-of-the-art results on the ArmGesture and Canal9 datasets. Notably, our approach achieves a near-perfect accuracy on the ArmGesture dataset (99.59%). For the NATOPS dataset, the state-of-the-art result is 87.00% by Song *et al.* [10]. Their approach used a multiview HCRF to jointly learn view-shared and view-specific hidden dynamics, where the two views are defined as upper body joint configuration and hand shape information. Even without considering the multi-view nature of the dataset (they perform an early-fusion of the two views), their approach achieved a comparable accuracy of 85.00%. This is still a significant improvement over various previous results using an early-fusion: HMM (from [10], 76.67%), HCRF (from [10], 76.00%), and HCNF (78.33%).

References

[1] Y. Bengio. Learning deep architectures for AI. *FTML*, 2(1):1–127, 2009. 1

[2] K. Bousmalis, L.-P. Morency, and M. Pantic. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *FG*, 2011. 2

[3] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, 2010. 1

[4] I. Laptev. On space-time interest points. *IJCV*, 64(2-3):107–123, 2005. 1

[5] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1

[6] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *CVPR*, 2007. 1

[7] J. Peng, L. Bo, and J. Xu. Conditional neural fields. In *NIPS*, 2009. 2

[8] A. Quattoni, S. B. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE TPAMI*, 29(10):1848–1853, 2007. 2

[9] Y. Song and R. Davis. Action recognition by hierarchical sequence summarization. In *CVPR*, 2013. 1, 2

[10] Y. Song, L.-P. Morency, and R. Davis. Multi-view latent variable discriminative models for action recognition. In *CVPR*, 2012. 2

[11] Y. Song, L.-P. Morency, and R. Davis. Multimodal human behavior analysis: Learning correlation and interaction across modalities. In *ICMI*, 2012. 2

[12] J. Sun, X. Wu, S. Yan, L. F. Cheong, T.-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *CVPR*, 2009. 1

[13] J. Wang, Z. Chen, and Y. Wu. Action recognition with multiscale spatio-temporal contexts. In *CVPR*, 2011. 1