Action Attribute Detection from Sports Videos with Contextual Constraints

Xiaodong Yu[1]
xiaodong_yu@cable.comcast.com
Ching Lik Teo[2]
cteo@cs.umd.edu
Yezhou Yang[2]
yzyang@cs.umd.edu
Cornelia Fermüller[2]
fer@cfar.umd.edu
Yiannis Aloimonos[2]
yiannisg@cs.umd.edu

[1] Comcast Corporation
Washington DC, USA
[2] University of Maryland,
College Park, MD, USA

In this paper, we study the problem of detecting action attributes from sport videos. Action attributes include atomic components of action classes (such as the motion patterns of human limbs and body), contextual components of action classes (such as the objects and scenes involved in the action), and non-semantic attributes, a.k.a data-driven attributes [2]. A common property of action attributes is that they can be generalized into different action classes. This is especially true in sports videos. For example, *bend*, as an action attribute that describes the motion of human body, is present in the action `tennis serve`, `bowling`, `snatch`, etc. That is why they can be learned even from training sets that contain only a few examples for each action class (*one-shot* learning) or even no example for some action classes (*zero-shot* learning). We focus on the action attributes related to motion patterns of human body in this paper; however, our model can be easily extended to detect the other types of action attributes as well.

The concept of action attribute was first introduced in [2]. However, the approach proposed in [2] has several severe limitations that restrict the applicability of action attributes. One of the most noticeable problems is that action attributes are labelled at the level of an action class, instead of being labelled at the level of a frame or at the level of a video. As a result, all videos from the same action class are assumed to have the same set of action attributes, regardless of the exact content of a specific video, and the temporal structure of the action attributes is totally discarded. But in reality, not every video belonging to the same action class have the same set of action attributes; more often than not, real-world videos would have some exceptions. For example, in a video of the `snatch` activity, the athlete may not be able to completely lift the barbell above his head at the end so we cannot say this video has an action attribute *two arms raise pose*. Furthermore, the temporal structure is a unique property of action attributes and we lose lots of descriptive capacity if we ignore it. For example, given a video of `basketball layup`, a description *"the athlete starts with a slow run and lasts for half a second, then jumps forward with single leg in the next second, finally jumps up and throws the ball (into the basket), and maintains a slow run at the end of the video"* will be more useful than simply saying *"there are slow running, jumping forward, jumping up, throwing in this video"*. The goal of this paper is thus to detect the key action attributes at each frame from a given video so that we can generate video descriptions at a much finer granularity than those from previous work.

While having great advantages as discussed above, it is obviously a much more challenging task to locate the temporal occurrences of every action attributes in a given video. As a high-level semantic concept, a particular action attribute may exhibit significant variability due to viewpoint changes, photometric measurements, intra-class variability (e.g. attribute *two arms open* can have different opening angles between two arms, *jump up* can have different height and velocity, etc). Naive detectors that rely entirely on local features will easily be overwhelmed by a large number of false positives and/or false negatives. So we must take into account the contextual constraints in both the temporal and semantic domains, thereby reducing the noise in the local feature space to produce more reliable results.

We take into account two types of contextual constraints in this paper: the temporal context and the semantic context. To promote agreement between the different attribute labels at different frames, we model the contextual constraints with a conditional random field (CRF) as shown in Figure 1. The features extracted from T frames in a video are denoted as a vector of T local observations, $\mathbf{X} = \{x_1, x_2, ..., x_T\}$; and at each local observation at frame t, by $\mathbf{x}_t$, the histogram of visual words as discussed in the preceding section. At each frame t, we wish to detect the presence of A action attributes $\mathbf{y}_t = \{y_t^1, y_t^2, ..., y_t^A\}$, which are the states in the CRF model. In the literature, this is also known as a factorial CRF [4]. In a factorial CRF, we have multiple states at each time point, $\mathbf{y}_t = \{y_t^1, y_t^2, ..., y_t^A\}$, and there are edges between every pair of y_t^i and y_t^j, $(i, j) \in \{1, 2, ..., A\}$[1]. To avoid clutter, we only show two attributes at each time point in Figure 1(b). In the experimental dataset used, there are 24 attributes at each time point. The between-chain edges are designed to promote agreement between different attributes at the same time point. The intuition is that some attributes tend to occur together, e.g.*two arms oscillate* and *fast run* while others don't, e.g. *slow run* and *fast run*. Thus the between-chain edges take into account the co-temporal correlation among attributes. The within-chain edges are designed to promote agreement between the states of adjacent time point based on the Markovian assumption and we enforce agreement between states in adjacent time points after accounting for the correlation between the neighboring temporal states. In this paper, the effectiveness of our methods are clearly illustrated by the experimental evaluations.

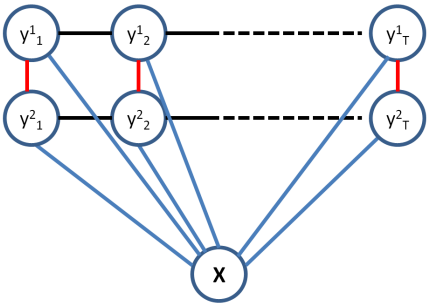

Figure 1: A factorial CRF model. To avoid clutter, we only show two attributes at each time points. In reality, we can have as many as attributes at each time points and they are fully connected to each other.

[1] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[2] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing Human Actions by Attributes. In *CVPR*, 2011.

[3] Cristian Sminchisescu, Atul Kanaujia, and Dimitris Metaxas. Conditional Models for Contextual Human Motion Recognition. *CVIU*, 104:210–220, 2006.

[4] Charles Sutton, Andrew McCallum, and Khashayar Rohanimanesh. Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. *Journal of Machine Learning Research*, 8:693–723, 2007.

[1] For clarity, we call the edges between states of the same time points as *between-chain edges* and the edges between states of adjacent time points as *within-chain edges*. In Figure 1, the former are colored in red and the latter in black