Weakly Supervised Action Detection

Parthipan Siva
psiva@eecs.qmul.ac.uk
Tao Xiang
txiang@eecs.qmul.ac.uk

School of EECS,
Queen Mary University of London,
London E1 4NS, UK

Detection of human action in videos has many applications such as video surveillance and content based video retrieval. Actions can be considered as spatio-temporal objects corresponding to spatio-temporal volumes in a video. The problem of action detection can thus be solved similarly to object detection in 2D images [3] where typically an object classifier is trained using positive and negative object examples and the detection is performed via 2D sliding window search. In our case, the classifier is applied with spatio-temporal subvolume search. The key problem, however, lies in training the spatio-temporal action volume classifier. Taking a conventional fully supervised approach, the spatio-temporal locations of the action of interest have to be manually annotated frame by frame in the training videos. This could be prohibitively expensive and subject to human bias. A data-driven automated annotation approach would be more desirable to deal with this ambiguity.

We propose to overcome the problem of manual action annotation of the training dataset by taking a weakly supervised learning (WSL) approach. Given a training dataset, the only annotation required by our WSL approach is the binary labelling of each video indicating whether the video contains the action of interest. More specifically, given a positive set of videos known to contain the action of interest and a negative set of videos without the action of interest, our WSL approach aims to determine automatically the spatial and temporal locations of the action in the positive set. We cast this WSL problem as a multiple instance learning (MIL) problem. Each video is considered as a bag of instances, i.e. candidate spatio-temporal volumes. A bag is either positive or negative depending on whether it contains positive instances (i.e. volumes containing an example of the target action). The objective of MIL is to identify the positive instances from the positive bags.

In this paper we make two main contributions: (1) To the best of our knowledge, this is the first weakly supervised action detection algorithm using only binary annotation of the training set. (2) We also present a novel global MIL technique that can localize the action of interest in a video with better results than the standard MIL techniques.

Proposed Method – Given a set of training videos, we define instances (potential locations for actions) as a spatio-temporal cuboids surrounding people detected by the pre-trained person detector of Felzenszwalb et al. [3]. These instances are then pruned and re-sampled to provide us with a final set of initial instances $\mathcal{C}_i^+ = \{c_{i,1}^+, c_{i,2}^+, \dots, c_{i,M}^+\}$ from the positive training videos $i = 1 \dots N^+$ and a set of negative instances $\mathcal{C}_i^- = \{c_{i,1}^-, c_{i,2}^-, \dots, c_{i,M}^-\}$ from the negative training videos $i = 1 \dots N^-$.

Now we want to select a set $\mathcal{G}^* = \{c_1, c_2 \dots, c_{N^+}\}$ consisting of one instance from each of the N^+ positive videos such that the selected instance is our action of interest. That is, $\mathcal{G}^*$ is the automatic annotation of the training set. We select this set globally using both inter- and intra-class measures by minimising the following cost function,

$$\mathcal{G}^* = \arg\min_{\mathcal{G}} \sum_{c_j \in \mathcal{G}} \left(D\left(c_j, \mathcal{G}_{-j}, k_p\right) + \left[1 - D\left(c_j, \mathcal{C}^-_{i=1 \dots N^-}, k_n\right)\right] \right) \quad (1)$$

where $\mathcal{G} = \{c_1, c_2, \dots, c_j, \dots, c_{N^+}\}$ is a set composed of one instance from each positive bag, $\mathcal{G}_{-j}$ is the set $\mathcal{G}$ excluding element c_j and $D(c, \mathcal{M}, k)$ defines the distance from a single instance c to a set of instances $\mathcal{M}$ with a constant parameter k (for the positive and negative training sets, it becomes k_p and k_n respectively). The term $D\left(c_j, \mathcal{G}_{-j}, k_p\right)$ aims to minimize the intra-class distances and the term $\left[1 - D\left(c_j, \mathcal{C}^-_{i=1 \dots N^-}, k_n\right)\right]$ is designed for maximizing the inter-class distances.

	Proposed Approach	*Siva et al. [5]	MI-SVM [1]	DD [4]	EMDD [7]
Boxing	40.7	**57.4**	20.4	9.3	24.1
Clapping	**79.4**	70.6	61.8	21.3	23.5
Handwaving	**93.6**	87.2	85.1	44.1	31.9

** Uses a single manual annotation.*

Table 1: Average annotation results (%) on training data.

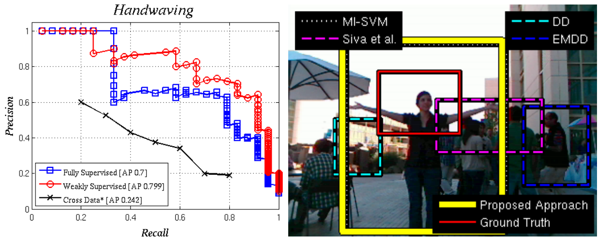

Figure 1: Handwaving detection result of WSL vs FSL on test data. WSL out performs FSL because the ground truth manual annotation of training data is biased. Our WSL approach (yellow) annotates the action including the hands where as the biased manual ground truth (red) does not.

In defining the distance function $D(c, \mathcal{M}, k)$ recall that each instance c is a potential action cuboid and as such is represented by a BoW histogram h_c. We thus define $d(c, m)$ the distance between two instances c and m as one minus the histogram intersection (HI) [6]. Then to compute $D(c, \mathcal{M}, k)$ we first sort all instances in $\mathcal{M}$ according to their distances to c in ascending order. Let each instance in this sorted set be m_l, then

$$D(c, \mathcal{M}, k) = \frac{1}{k} \sum_{l=1}^{k} d(c, m_l). \quad (2)$$

We are taking the average distance from instance c to the closest k instances in set $\mathcal{M}$.

The cost function, Eq. 1, is minimized using a genetic algorithm and the resulting automatic annotation $\mathcal{G}^*$ is used to train a SVM as the final detector.

Results – We test our approach on the 3 action classes found in the MSR2 dataset [2]. The automatic annotation of the training data (set $\mathcal{G}^*$) is compared with existing MIL techniques MI-SVM [1], DD [4], EM-DD [7] and the approach of Siva et al. [5] which uses one manual annotation. Our approach performs significantly better than the existing MIL techniques (MI-SVM, DD and EM-DD) as shown in Table 1.

We compare the detector trained by the weakly supervised approach to the detector trained by the fully supervised approach, which uses full manual annotation of the training data. We find that the weakly supervised approach performs comparably to the fully supervised approach and in fact for the handwaving class the weakly supervised detector outperforms the fully supervised approach. This can be attributed to a bias in the manual annotation of the training data. As seen in Fig. 1 the manual annotation of some sequences does not include the forearm and hands where as the automatic annotation does include the forearm and hands which is in motion during the action.

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*, pages 561–568. MIT Press, 2003.

[2] L. Cao, Z. Liu, and T. S. Huang. Cross-data action detection. In *CVPR*, 2010.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 99(9):1627–1645, 2009.

[4] O. Maron and T. Lozano-Perez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 1998.

[5] P. Siva and T. Xiang. Action detection in crowds. In *BMVC*, 2010.

[6] M. Swain and D. Ballard. Color indexing. *IJCV*, 7(1):11–32, 1991.

[7] Q. Zhang and S. A. Goldman. Em-dd an improved multiple-instance learning technique. In *NIPS*, 2001.