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

Abstract

The detection of human action in videos of busy natural scenes with dynamic background is of interest for applications such as video surveillance. 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, which is tedious and unreliable. In this paper, for the first time, a weakly supervised action detection method is proposed which only requires binary labels of the videos indicating the presence of the action of interest. Given a training set of binary labelled videos, the weakly supervised learning (WSL) problem is recast as a multiple instance learning (MIL) problem. A novel MIL algorithm is developed which differs from the existing MIL algorithms in that it locates the action of interest spatially and temporally by globally optimising both inter- and intra-class distance. We demonstrate through experiments that our WSL approach can achieve comparable detection performance to a fully supervised learning approach, and that the proposed MIL algorithm significantly outperforms the existing ones.

1 Introduction

Detection of human action in videos has many applications such as video surveillance and content based video retrieval. Action detection [7, 8, 14, 17] is different from the extensively studied action recognition problem [10, 12, 16]. In action recognition, one assumes that each video has been pre-segmented to contain only a complete action sequence. The task is to classify the whole action video into one of the known action categories. In contrast, action detection aims to both recognise the action and estimate where it occurs spatially and temporally in a video that may contain multiple background actions. Detection is therefore a much harder problem than recognition. Detection is needed because real world actions in a public space typically occur in a crowded and dynamic environment.

Actions can be considered as spatio-temporal objects corresponding to spatio-temporal volumes in a video [17]. The problem of action detection can thus be solved similar to object detection in 2D images [5] 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. Importantly, manual annotation is subjective and can thus be biased and suboptimal.

BMVC 2011 http://dx.doi.org/10.5244/C.25.65

For instance, different people may have different understandings of what defines a handwaving action. Does it involve only hands, or should it include arms, upper body, or even the whole body? 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 present a novel MIL algorithm which differs from the existing MIL algorithms in that it locates the action of interest spatially and temporally by optimising both inter- and intra-class distance of the globally selected positive instances. Our experiments on a public dataset demonstrate that a detector learned using our approach can achieve comparable performance to the fully supervised approach. We also show that our new MIL approach can localize actions of interest in the training set with a significantly greater accuracy than the existing alternative MIL techniques.

1.1 Related Work

Due to the prohibitive cost of manual labelling of a training video set, most existing work action detection avoids a fully supervised approach. Earlier methods fall into two categories: single example query [8, 17] and cross dataset training [2, 18]. In the single example query approach [8, 17], one example of the action of interest is manually annotated as a template or query sample. Using this single example, test videos are queried for the action of interest. This method cannot handle the intra-class variation of the actions caused by different people performing the actions at different camera viewpoints. In the cross dataset training approach [2, 18] the actions are learned using a clean positive training set (e.g. those used for action recognition) captured in a different environment as the test videos (hence cross dataset). In this clean training set, the action is pre-segmented and performed in the absence of background activity, so manual annotation of spatial and temporal location is not needed. While this approach can handle intra-class variations, obtaining a clean positive training set in the absence of background activity is not always feasible. Effectively transferring the learned action detector from one dataset to another is also far from being solved.

Recently, Hu et al. [7] and Siva et al. [14] have attempted weakly supervised learning for action detection. However, both methods rely on more manual annotation efforts than ours. Similar to our method, Hu et al. [7] also adopt a MIL method based on that of Andrew et al. [1]. However, in addition to the binary label for each training video, their method requires the manual annotation of an approximate spatial and temporal location of the head of the person performing the action. By doing so, all the background actions are eliminated. These background actions can be potentially confused with the target action thus causing problems for MIL. The removal of them makes the Hu et al. [7] problem much easier to solve than our MIL problem. In addition, although the amount of annotation is reduced compared to a fully supervised approach, it is still substantial. In the method of Siva et al. [14], in additional to the binary labels, a single manually annotated action cuboid/volume needs to be provided. Using this cuboid, a greedy k nearest neighbour (kNN) search is performed on the positive training set to obtain the spatial and temporal location of the action of interest. Compared to their method, our approach does not require the manually annotated action example. Furthermore, their iterative annotation process only finds action examples that are similar to the manually annotated one. It can thus only handle small intra-class variations. In contrast, our method uses a global optimization process that is capable of handling larger intra-class variations.

Multiple Instance Learning (MIL) was first introduced by Dietterich et al. [4] for the problem of drug activity prediction. For MIL, data is represented as bags and each bag contains a set of instances. A positive bag contains at least one positive instance and a negative bag contains no positive instances. For our problem each video clip in the training set is considered as a bag. Instances are spatio-temporal cuboids of potential action locations in the video and a positive instance contains the action of interest. The task is to find the correct positive instances in the positive training bags. There have been many MIL instance classifiers proposed in the past including DD [11], EM-DD [19] and MI-SVM [1], and their variants. However, these methods either iteratively select positive instances locally [11, 19] (that is, each positive instance is selected independently of each other), or select the positive instances globally (by considering only the distances between positive and negative instances). In this work, we present a global method for MIL that exploits both the positive instance compactness (intra-class distances) and distances from negative instances (inter-class distances). Recently Deselaers et al. [3] presented an alternative global instance selection method based on conditional random field that considers both intra- and inter-class distance. However, unlike our method, theirs has a complex formulation with many parameters that must be tuned on an auxiliary dataset.

To summarize, the main contributions of the paper are: (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 existing MIL techniques.

2 Proposed Approach

Our goal is to train an action detector with weakly labelled data, i.e. a positive set of videos known to contain at least one occurrence of the action of interest and a negative set of videos known to contain no action of interest. From this our algorithm automatically annotates the action of interest in the positive set of videos. Using the automatic annotation an action detector can be trained. The detector is then used in a sliding window fashion (spatio-temporal volume search) to detect the occurrence of the action in the test videos.

Before proceeding to the automatic annotation algorithm we first give a brief overview of how we represent an action. An action is represented as a spatio-temporal cuboid/volume in a video. We will refer to this as the action cuboid. The action cuboid is described using a bag of words (BoW) histogram. We use the spatio-temporal interest point (STIP) descriptors of Laptev et al. [9] as features. To create the BoW representation of the action cuboids, we cluster 100000 randomly selected STIPs from the training dataset into 2000 code words using k-means clustering. All action cuboids are then represented as a 2000 bin histogram of the STIPs inside the action cuboid.

2.1 Automated Annotation of Training Set

Taking a MIL approach, we consider each video, in both the positive and negative set, as a bag and in each bag we define a set of instances which are the potential action cuboids in that video. Once the instances are defined we can select one instance from each positive bag as the action of interest.

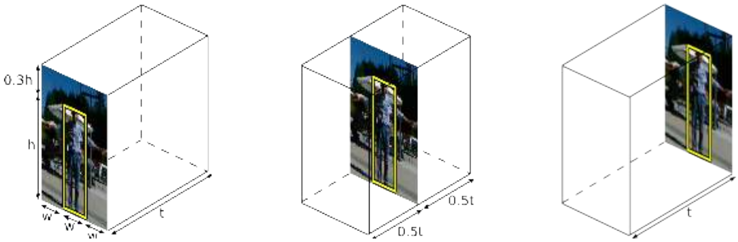


Figure 1: Given a detected person bounding box w by h on frame f the action cuboids of temporal length t at different temporal location, relative to frame f, are constructed.

2.1.1 Instance Definition

For each video in both the positive and negative training sets we need to define instances as potential action cuboids. A straightforward way is to define instances as all possible cuboids of different sizes that can fit within the video. A video sequence of 160x120 lasting one minute at 25FPS contain over 1 billion valid action cuboids, making any MIL algorithm intractable. Therefore, we have to screen the cuboids to limit the number of feasible instances.

Since we are only interested in actions being performed by stationary people, we create an initial set of instances $\mathcal{C}'$ surrounding people detected by a state-of-the-art person detector [5]. The person detector is run on every F^{th} frame and at each detected person location a set of action cuboids are created. For a detected person of height h and width w, the corresponding action cuboid has a spatial size $3w$ by $1.3h$ (Fig. 1). The action cuboid is larger than the detected person size because the person detector is trained to detect people at a neutral pose without outstretched arms or legs. By including a buffer of w pixels on both the left and right sides of the detected person and a buffer of $0.3h$ above the head, we can account for extension, of the hands and legs during various actions. We consider multiple temporal sizes, $t \in \{t_k\}$ for the action cuboid as we do not know the duration of the action of interest. Three different temporal locations of the action cuboid are considered relative to the frame in which the person was detected, as illustrated in Fig. 1. This is to account for missed person detection on some frames during an action.

The initial set $\mathcal{C}'$ still numbers in the thousands of action cuboids and can be further pruned to a more compact and reliable cuboid set. We first rank the cuboids in $\mathcal{C}'$ based on STIP density and temporal spread using Algorithm 1. From the ranked list of cuboids $\mathcal{C}''$ we select the first M cuboids as the reliable cuboid set $\mathcal{C}$ for use as instances. In this way we can eliminate false positives by the person detector on static background and stationary people as they will not produce STIPs. Note that background/negative action instances can also produce dense STIP points (in some case denser than the positive ones). One is thus in danger of removing the potential positive instance by relying on STIP density alone for ranking. To overcome this problem, we remove and reintroduce the STIPs during the ranking process (Algorithm 1, Lines 8 and 11). These steps are important to ensure that the instances in $\mathcal{C}$ contains samples from the entire video. Without these steps, $\mathcal{C}$ could contain many overlapping samples from a single high STIP density location in the video which may not contain the action of interest.

```
input  : $\mathcal{C}'$ – Action cuboids surrounding each person detection.
         $S$ – List of $(x,y,t)$ location of all STIPs.
         $T$ – minimum action cuboid STIP density.
output : $\mathcal{C}''$ – Ranked list of action cuboids.
$\mathcal{C}'' = \{\}, \mathcal{C}_r = \mathcal{C}', S_r = S$ ;
while $S_r \neq \{\}$ do
    $d_{max}$ = highest density of cuboid in $\mathcal{C}_r$ where density $= \frac{\#STIP}{\text{cuboid volume}}$ ;
    if $d_{max} > T$ then
        $c_{max}$ = cuboid in $\mathcal{C}_r$ with highest STIP density ;
        $S_{max}$ = all STIP points inside $c_{max}$ ;
        Remove $c_{max}$ from $\mathcal{C}_r$ ;
        Remove $S_{max}$ from $S_r$ ;
        Add $c_{max}$ to $\mathcal{C}''$ ;
    else
        $S_r = S$ ;
    end
end
```

Algorithm 1: Rank action cuboids

2.1.2 Positive Instance Selection

Having defined instances in a video, we now have a set of instances $\mathcal{C}_i^+ = \{c_{i,1}^+, c_{i,2}^+, \ldots, c_{i,M}^+\}$ from the positive videos $i = 1 \ldots N^+$ and a set of negative instances $\mathcal{C}_i^- = \{c_{i,1}^-, c_{i,2}^-, \ldots, c_{i,M}^-\}$ from the negative videos $i = 1 \ldots N^-$. We want to select a set $\mathcal{G}^* = \{c_1, c_2 \ldots, c_{N^+}\}$ consisting of one instance from each of the N^+ positive videos such that the selected instance is our action of interest. We select this set globally using both inter- and intra-class measures. Specifically $\mathcal{G}^*$ is selected by minimising the following cost function,

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

where $\mathcal{G} = \{c_1, c_2, \ldots, c_j, \ldots, c_{N^+}\}$ 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 $D(c_j, \mathcal{G}_{-j}, k_p)$ aims to minimize the distance between the instances selected from each positive videos, that is to minimize the intra-class distance to ensure that the selected instances look similar to each other. The term $\left[1 - D\left(c_j, \mathcal{C}^-_{i=1\ldots N^-}, k_n\right)\right]$ is designed for maximizing the distance between the instances selected from each positive videos and the instances in all the negative videos. By maximizing the inter-class distance we can ensure that the selected instances look dissimilar to those in the negative videos.

Distance Metric – We need to define a distance metric $D(c, \mathcal{M}, k)$ between a single instance and set of instances taking into account that the set $\mathcal{M}$ can be multi-modal (e.g. caused by variations in viewpoint) and noisy. Recall that each instance c is a potential action cuboid and as such is represented by a BoW histogram h_c. We thus define the distance between two instances c and m as one minus the histogram intersection (HI) [15].

$$d(c,m) = 1 - \sum_{i=1}^{2000} min(h_c(i), h_m(i)) \tag{2}$$

where $h_c(i)$ and $h_m(i)$ are the i^{th} bin of the normalized BoW histogram representation of instance c and m. Now the value of $d(c,m)$ ranges from 0 to 1. It assumes the value 0 when instances c and m are identical to each other and 1 when they look completely different.

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

$$D(c,\mathcal{M},k) = \frac{1}{k}\sum_{l=1}^{k} d(c,m_l). \tag{3}$$

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

Optimization Method – To solve Eq. (1), we use the genetic algorithm (GA) [6] implementation in MATLAB. A GA is an evolutionary algorithm that selects the optimal solution using techniques inspired by evolution. A population of random candidate solutions $(\mathcal{G}_1,\ldots,\mathcal{G}_n)$ evolves through reproduction and random mutation towards the optimal solution $(\mathcal{G}^*)$. In our reproduction step a child $\mathcal{G}^{\text{child}}$ is created from parents $\mathcal{G}^{\text{P1}}$ and $\mathcal{G}^{\text{P2}}$ as follows:

$$\mathcal{G}^{\text{P1}} = \{c_1^{\text{P1}},\ldots,c_m^{\text{P1}},c_{m+1}^{\text{P1}},\ldots,c_{N^+}^{\text{P1}}\} \quad \mathcal{G}^{\text{P2}} = \{c_1^{\text{P2}},\ldots,c_m^{\text{P2}},c_{m+1}^{\text{P2}},\ldots,c_{N^+}^{\text{P2}}\}$$
$$\mathcal{G}^{\text{child}} = \{c_1^{\text{P1}},\ldots,c_m^{\text{P1}},c_{m+1}^{\text{P2}},\ldots,c_{N^+}^{\text{P2}}\}$$

where m is randomly selected. Mutation occurs by randomly switching instances from a bag.

2.2 Detector

Given a set of selected positive action cuboids $\mathcal{G}^*$ and a set of videos without the action of interest we train a support vector machine (SVM) as our action cuboid classifier. Since the number of positive action cuboids, N^+, is much smaller than the potential set of negative cuboids, we employ the positive mining technique of Felzenszwalb et al. [5]. We use the histogram intersection kernel [15] for the support vector machine.

For both the negative instance mining and detection in test videos we fix the aspect ratio of the search window based on the aspect ratio of the cuboids in $\mathcal{G}^*$. These fixed aspect ratios are related to the aspect ratios of each of the component in the person detector [5]. The temporal duration of the search window is fixed to the same values $t \in \{t_k\}$ used in defining the action cuboids in Section 2.1.1.

3 Experiments

Datasets – Experiments were carried out using the MSR2 dataset [2] which is the biggest action detection dataset publicly available. The MSR2 dataset contains 54 videos with three action categories: boxing, clapping and handwaving (see Fig. 2). Each video contains at least three action separated temporally. We split each of the 54 videos to contain only one action; the split occurs at the midpoint between the end of the last action and the start of the

	Proposed Approach	*Siva et al. [14]	MI-SVM [1]	DD [11]	EMDD [19]
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 (%).

next action. If multiple actions overlap temporally they are all included in one clip. Note this split is different from temporal segmenting the actions because the action can still start and finish at variable temporal locations in each split video. After splitting the videos there are 181 videos of which 16 contains multiple actions. A two-to-one random division of the 181 videos is used as the training and testing set. During the division, the 16 videos containing multiple actions are always included in the testing set. Localization of actions in the training set is necessary because, among the detected STIPs, on average only 19% belong to the action of interest while the rest come from the background actions. The spatio-temporal location of each action in each video is provided together with the data. They are not used for training, but as ground truth for performance evaluation.

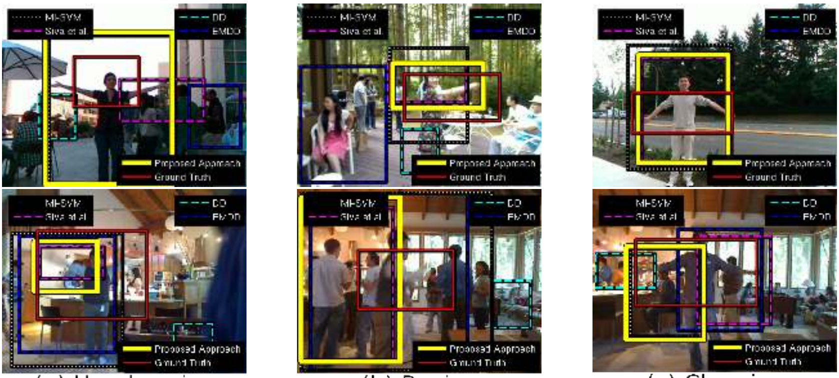
Competitors – For the automated annotation of the training data, we compare our global optimization solution to the widely used MI-SVM approach of Andrew et al. [1] which exploits inter-class distance globally. We also compare with local intra-class distance based MIL algorithms DD [11] and EM-DD [19]. In addition, we compare our automated annotation to the approach of Siva et al. [14] where a single video clip from the training set is randomly selected and manually annotated. For a fair comparison, we re-implement Siva et al.'s method using the same action representation based on STIPs, instead of the track features, and histogram intersection distance, rather than the chi-squared distance. For detection result on the testing set we compare a detector trained with our weakly supervised annotation to a detector trained with manual annotation, i.e. a fully supervised learning approach.

Settings – For instance definition (Section 2.1.1), we run the pre-trained person detector provided by Felzenszwalb et al. [5] at a frame rate of 5FPS and consider action cuboid temporal durations of $t_k \in \{75, 100\}$ frames. During the pruning stage (Algorithm 1), a minimum action cuboid STIP density threshold of $T = 0.0002$ was used. In practice, any value very close to zero will make little difference here. For instance definition, the cuboids in each video are screened to $M = 200$ instances per bag. For instance selection (Section 2.1.2), we need to set the two parameters k_p and k_n (see Eq. (1)). k_p and k_n are based on the number of positive bags (N^+) and the number of negative instances respectively. We found that the result is stable when k_p is in the range of 20% to 70% of N^+ and k_n is in the range of 5% to 25% of M (number of instances per bag). In our experiments, we used $k_p = 25$ and $k_n = 10$ for all classes. Finally, for the genetic algorithm a population size of 2000 and a mutate chance of 10% was used.

3.1 Automated Annotation Results

To evaluate the effectiveness of the proposed MIL algorithm for automated action annotation, we calculate the percentage of correctly detected actions in the training data. In accordance with [18] and [2] we define detection as correct if at least 1/8 of the volume size overlaps with the ground truth. The detection results are summarized in Table 1 and some examples of the automated annotation result can be seen in Fig. 2.

The results show that the proposed MIL algorithm based on global inter- and intra-class distance optimization achieves a higher correct annotation rate for all action categories than



(a) Handwaving (b) Boxing (c) Clapping

Figure 2: Examples of automated annotation of training data using different MIL algorithms.

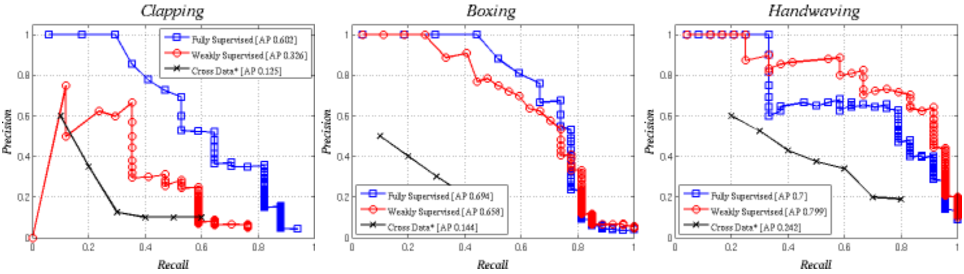


Figure 3: Test data detection precision recall curve. Cross dataset results as published in [2].

the standard MI-SVM algorithm. In particular for the boxing class the annotation accuracy is 2 times better. The MI-SVM considers only the separability of the positive instances from the negative instances (inter-class distance). The results show that in the absence of ground truth, considering the compactness of the positive instances as well as the separability of positive and negative instance is more effective in selecting the correct positive instances. Similarly, our algorithm significantly outperforms the local intra-class distance based algorithms: DD and EM-DD. For action detection, due to the large number of negative background activities, a global intra-class measure of the selected positive instances is crucial. It can also be seen that our algorithm achieves better performance on two of the three action categories, compared with the method of Siva et al. [14] even though it uses only a single manual annotation. The advantage of our algorithm is particularly clear for clapping. This is because people in the dataset performed clapping in quite different ways (see Fig. 2(c)) - some at shoulder height, others at waist height - resulting in large intra-class variation which the method of Siva et al. cannot cope with.

3.2 Detection Results

We compare the detection performance of the weakly supervised detector with a fully supervised detector using the precision recall curve (PRC) as defined in [18]. The PRCs are

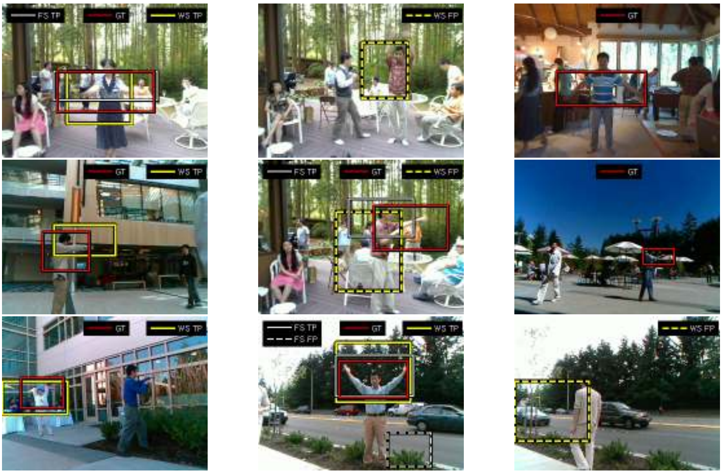


Figure 4: Detection examples of fully supervised (FS) and weakly supervised (WS) detectors on the testing data. In the images TP - true positive, FP - false positive and GT - ground truth.

presented in Fig. 3. Example detection on testing data is presented in Fig. 4.

It can be seen that the weakly supervised detector achieves a higher average precision (AP) than the fully supervised detector on detecting handwaving (AP of 0.799 vs 0.700). This can be attributed to a bias in the manual annotations. In some videos the manual annotation cuboid provided with the dataset does not encompass the entire extent of the hand motion (Fig. 2(a)) and as such does not include all the useful STIPs that occur during the action. Our automatically annotated cuboids include the entire hand motion range and thus include more relevant information for the detector to learn.

The weakly supervised detector is able to achieve similar performance as the fully supervised detector on boxing. The weakly supervised clapping detector has the worst performance. However, it is still able to obtain an average precision that is about 50% of the fully supervised detector. There are two possible reasons for the poor performance of the weakly supervised clapping detector: 1) the clapping class has fewer training samples than boxing and handwaving (33 videos contain clapping as opposed to 47 and 53 for boxing and handwaving). MIL algorithms in general struggle with few training samples. 2) Clapping is a highly symmetric action. For a MIL algorithm, the movements of either left or right hand are equally plausible as the two hand movements for defining clapping because all of them always appear in each positive video. As a result, the automated annotation of clapping cuboids often contain only one hand movements, resulting in low detection accuracy. This is an intrinsic problem for a WSL approach that uses only binary labels. Unless one hand 'clapping' is part of the negative videos, this problem cannot be addressed.

In Fig. 3 we also plot the cross dataset detection results as reported in [2]. The clean KTH dataset [13] is used for training a detector which is then adapted to the MSR2 dataset for detection. While this is not directly comparable with our method, as part of the cross dataset detection test set was used as our training set, it does indicate the weak performance in using a different training set. This is despite the fact that the clean KTH dataset contains no background action and videos are pre-segmented.

4 Conclusion

We have presented a weakly supervised approach to action detection that, unlike existing methods, requires no manual annotation other than a binary label indicating the presence of the target action in a video. The key component of the approach is a novel multiple instance learning (MIL) algorithm that exploits both inter- and intra-class distances globally. Our experiments demonstrate the superior performance of the proposed MIL algorithm compared with a number of existing MIL algorithms. Most encouraging of all, we show that in some cases, the weakly supervised detector can even outperform a fully supervised detector by avoiding the inaccuracy and bias in human manual annotation.

References

[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] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *ECCV*, 2010.

[4] T. G. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 1997.

[5] 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.

[6] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1989. ISBN 0201157675.

[7] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T.S. Huang. Action Detection in Complex Scenes with Spatial and Temporal Ambiguities. In *ICCV*, 2009.

[8] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV*, 2007.

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

[10] A. P. B. Lopes, E. A. Jr. Valle, J. M. Almeida, and A. A. Araújo. Action recognition in videos: from motion capture labs to the web. *CoRR*, abs/1006.3506, 2010.

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

[12] R. Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990, 2010.

[13] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, 2004.

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

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

[16] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding*, 115(2):224 – 241, 2011.

[17] W. Yang, Y. Wang, and G. Mori. Efficient human action detection using a transferable distance function. In *ACCV*, 2009.

[18] J. Yuan, Z. Liu, and Y. Wu. Discriminative video pattern search for efficient action detection. *TPAMI*, 2011.

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