Discovering Multi-Label Actor-Action Association in a Weakly Supervised Setting

Sovan Biswas[1] and Juergen Gall[1]

University of Bonn, 53115 Bonn, Germany
{biswas,gall}@iai.uni-bonn.de

Abstract. Since collecting and annotating data for spatio-temporal action detection is very expensive, there is a need to learn approaches with less supervision. Weakly supervised approaches do not require any bounding box annotations and can be trained only from labels that indicate whether an action occurs in a video clip. Current approaches, however, cannot handle the case when there are multiple persons in a video that perform multiple actions at the same time. In this work, we address this very challenging task for the first time. We propose a baseline based on multi-instance and multi-label learning. Furthermore, we propose a novel approach that uses sets of actions as representation instead of modeling individual action classes. Since computing the probabilities for the full power set becomes intractable as the number of action classes increases, we assign an action set to each detected person under the constraint that the assignment is consistent with the annotation of the video clip. We evaluate the proposed approach on the challenging AVA dataset where the proposed approach outperforms the MIML baseline and is competitive to fully supervised approaches.

1 Introduction

In recent years, we have seen a major progress for spatially and temporally detecting actions in videos [1–10]. For this task, the bounding box of each person and their corresponding action labels need to be estimated for each frame as shown in Figure 1. Such approaches, however, require the same type of dense annotations for training. Thus, collecting and annotating datasets for spatio-temporal action detection becomes very expensive.

To alleviate this problem, weakly supervised approaches have been proposed [11–13] where the bounding boxes are not given, but only the action that occurs in a video clip. Despite the promising results of the weakly supervised approaches for spatio-temporal action detection, current approaches are limited to video clips that predominantly contain a single actor performing a single action as in the datasets UCF 101 [14] and JHMDB [15]. However, most real world videos are more complex and contain multiple actors performing multiple actions simultaneously. In this paper, we move a step forward and introduce the task of weakly supervised multi-label spatio-temporal action detection with multiple actors in a video. The goal is to infer a list of multiple actions for each actor in

Fig. 1. The image shows a scene where two persons are talking. In this case there are two person that perform multiple actions at the same time. **Person A** indicated by the **blue** bounding box performs the actions *Stand*, *Listen to*, and *Watch*. **Person B** indicated by the **orange** bounding box performs the actions *Stand*, *Talk to*, and *Watch*. While in the supervised setting this information is also given for training, we study for the first time a weakly supervised setting where the video clip is only annotated by the actions *Stand*, *Listen to*, *Talk to*, and *Watch* without any bounding boxes or associations to the present persons.

a given video clip as in the fully supervised case [5–10]. However, in the weakly supervised setting only actions occurring in each training video are known. Any spatio-temporal information about the persons performing these actions is not provided. This is illustrated in Figure 1 that shows two people standing and chatting. The video clip is only annotated by the four occurring actions *Stand*, *Listen to*, *Talk to*, and *Watch*. Additional information like bounding boxes or the number of present persons is not provided. In contrast to previous experimental settings for weakly supervised learning, the proposed task is much more challenging since a video clip can contain multiple persons, each person can perform multiple actions at the same time, and multiple persons can perform the same action. For instance, both persons in Figure 1 perform the actions *Stand* and *Watch* at the same time.

In order to address multi-label spatio-temporal action detection in the proposed weakly supervised setup, we first introduce a baseline that uses multi-instance and multi-label (MIML) learning [16–18]. Second, we introduce a novel approach that is better suited for the multi-label setting. Instead of modeling the class probabilities for each action class, we build the power set of all possible action combinations and model the probability for each subset of actions. Using a set representation has the advantage that we model directly the combination of multiple occurring actions instead of the probabilities of single actions. Since computing the probabilities for the full power set becomes intractable as the number of action classes increases, we assign an action set to each detected person under the constraint that the assignment is consistent with the annotation of the video clip. This is done by linear programming, which maximizes the overall gain across all plausible actors and action subset combinations. We evaluate the

proposed approach on the challenging AVA 2.2 dataset [19], which is currently the only dataset that can be used for evaluating this task. In our experiments, we show that the proposed approach outperforms the MIML baseline by a large margin and that the proposed approach achieves 83% of the mAP compared to a model trained with full supervision.

In summary, the contribution of this paper is three-fold:

- We introduce the novel task of weakly supervised multi-label spatio-temporal action detection with multiple actors.
- We introduce a first baseline for this task based on multi-instance and multi-label learning.
- We propose a novel approach based on an action set representation.

2 Related Work

Spatio-Temporal Action Detection. A popular approach for fully supervised spatio-temporal action detection comprises the joint detection and linking of bounding boxes [1, 3, 4, 20]. These linked bounding boxes form tubelets which are subsequently classified. Recently, many methods [10, 9, 21, 22] use standard person detectors for actor localization and focus on learning implicitly or explicitly spatio-temporal interactions. All these approaches, however, require that each frame is annotated with person locations and corresponding action labels. Since such dense annotations are expensive to obtain on a large scale, recent approaches [19, 23, 8] deal with temporally sparse annotations. Here, the action labels and locations are annotated only for a subset of frames. Even though there is a reduction in annotation, these methods still require person specific bounding boxes and their actions. Very few methods such as [11, 13] explore the possibility of weakly supervised learning. Most of these methods such as [24, 25] use multiple instance learning to recognize distinct action characteristics. These works, however, consider the case where a single person performs not more than one action.

Actor-Action Associations. Actor-action associations have been key to identify actions both in a fully supervised and weakly supervised settings. [26] performs soft actor-action association using tags as pre-training on a very large dataset for fully supervised action recognition. With respect to weak supervision, a few approaches use movie subtitles [27, 28] or transcripts [29, 30] to temporally align actions to frames. In terms of actor-action associations for multiple persons, [31, 32] associate a single action to various persons. To the best of our knowledge, our work is the first to perform multi-person and multi-label associations.

Multi-Instance and Multi-Label Learning. In the past, many MIML algorithms [33, 34] have been proposed. For example, [17] propose the MIMLBoost and MIMLSVM algorithms based on boosting or SVMs. [35] optimize a regularized rank-loss objective. MIML has been also used for different computer vision applications such as scene classification [16], multi-object recognition [18], and image tagging [36]. Recently, MIML based approaches have been used for action recognition [32, 37].

3 Multi-Label Action Detection and Recognition

Given a video clip with multiple actors where each actor can perform multiple actions at the same time as shown in Figure 1, the goal is to localize these actors and predict for each actor the corresponding actions. In contrast to fully supervised learning, where bounding boxes with multiple action labels are given for training, we address for the first time a weakly supervised setting where only a list of actions is provided for each video clip during training. This is a very challenging task as we do not know how many actors are present and each actor can perform multiple actions at the same time. This is in contrast to weakly supervised spatio-temporal action localization where it is assumed that only one person is in the video and that the person does not perform more than one action at a given point in time.

In order to address this problem, we first discuss a baseline, which uses multi-instance and multi-label (MIML) learning [16–18], in Section 4. In Section 5, we will then propose a novel method which uses a set representation instead of a representation of individual actions. This means that we build from the annotation of a video clip the power set of all possible action combinations. For example, the power set Ω for the three action labels *Listen*, *Talk*, and *Watch* is given by $\{\varnothing, \{Listen\}, \{Talk\}, \{Watch\}, \{Listen, Talk\}, \{Listen, Watch\}, \{Talk, Watch\}, \{Listen, Talk, Watch\}\}$. We then assign one set $\omega_i \in \Omega \setminus \varnothing$ to each actor a_i under the constraint that each action c occurs at least once, i.e., $c \in \bigcup_i \omega_i$. Using a set representation has the advantage that we model directly the combination of multiple occurring actions instead of the probabilities of single actions.

4 Multi-instance and Multi-label (MIML) Learning

One way to address the weakly supervised learning problem is to use multiple-instance learning. Since we have a multi-label problem, i.e., an actor can perform multiple actions at the same time, we use the concept of multi-instance and multi-label (MIML) learning [16–18]. We first use a person detector [38] to spatially localize the actors in a frame t and use a 3D-CNN such as I3D [39] or Slowfast [10] for predicting the action probabilities similar to fully supervised methods [8, 9]. However, we use the MIML loss to train the networks.

We denote by $A_t = \{a_1^t, a_2^t, \ldots, a_{n_t}^t\}$ the detected bounding boxes and by $f(a_i^t)$ the class probabilities that are predicted by the 3D-CNN. Let Y be the vector which contains the annotations of the video clip, i.e., $Y(c) = 1$ if the action class c occurs in the video clip and $Y(c) = 0$ otherwise. In other words, the bag A_t is labeled by $Y(c) = 1$ if at least one actor performs the action c and by $Y(c) = 0$ if none of the actors performs the action. The MIML loss is then given by

$$\mathcal{L}_{MIML} = \mathcal{L}\left(Y, \max_i(f(a_i^t))\right) \tag{1}$$

where $\mathcal{L}$ is the binary cross entropy. This means that the class probability should be close to one for at least one bounding box if the action is present and it should be close to zero for all bounding boxes if the action class is not present.

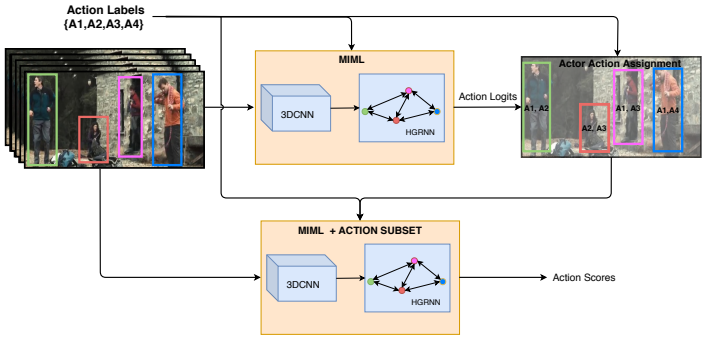


Fig. 2. Overview of the proposed approach. Given a training video clip with action labels {A1,A2,A3,A4}, we first detect persons in the video. We then train a 3D CNN with a graph RNN that models the spatio-temporal relations between the detected persons using the MIML loss to obtain initial estimates of the action logits. During actor-action association, subsets of the action labels are assigned to each detected person. The training of the network is continued using the MIML loss and the actor-action associations.

5 Actor-Action Association

While multi-instance and multi-label learning discussed in Section 4 already provides a good baseline for the new task of weakly supervised multi-label action detection, we propose in this section a novel method that outperforms the baseline by a large margin. As discussed in Section 3, the main idea is to change the representation from individual action labels to sets of actions. This means that we have one probability for a subset of actions $\omega \in \Omega$ instead of C probabilities where C is the number of action labels. We discuss how the probability of a set actions is estimated in Section 5.1. Due to the weakly supervised setting not all combinations of subsets are possible for each video clip. We therefore assign an action set $\omega \in \Omega$ to each actor a under the constraint that the assignment is consistent with the annotation of the video clip, i.e., each annotated action c needs to occur at least once and actions that are not annotated should not occur. The assignment is discussed in Section 5.2.

Figure 2 illustrates the complete approach. As described in Section 4, we use a 3D CNN such as I3D [39] or Slowfast [10]. Since the actors in a frame often interact with each other, we use a graph to model the relations between the actors. The graph connects all actors and we use a graph RNN to infer the action probabilities for each actor based on the spatial and temporal context. In our approach, we use the hierarchical Graph RNN (HGRNN) [7] where the features per node are obtained by ROI pooling over the 3D CNN feature maps. The HGRNN and 3D CNN are learned using the MIML loss (1). From the action class probabilities, we infer the action set probabilities as described in Section 5.1 and we infer the action set for each actor as described in Section 5.2. Finally,

we train the HGRNN and the 3D CNN based on the assignments. This will be discussed in Section 5.3.

5.1 Power Set of Actions

In principle, we could modify our network to predict the probability for each subset of all action classes instead of the probabilities for all action classes. However, this is infeasible since the power set of all actions is very large. If C is the number of actions in a dataset, the power set for all actions consists of 2^C subsets. Already with 50 action classes, we would need to predict the probabilities for over one quadrillion subsets. Instead, we use an idea that was proposed for HEX graphs [40] where the probabilities of a hierarchy are computed from the probabilities of the leave nodes. While we do not use a hierarchy, we can compute the probability of a subset of actions from the predictions of a network for individual actions.

Let $s_c \in (-\infty, \infty)$ denote the logit that is predicted by the network for the action class c. The probability of a subset of actions ω can then be computed by

$$p_\omega = \frac{\exp\left(\sum_{c\in\omega} s_c\right)}{\sum_{\omega'} \exp\left(\sum_{c\in\omega'} s_c\right)}. \tag{2}$$

The normalization term, however, is still infeasible to compute since we still need to sum over all possible subsets (ω') for the dataset.

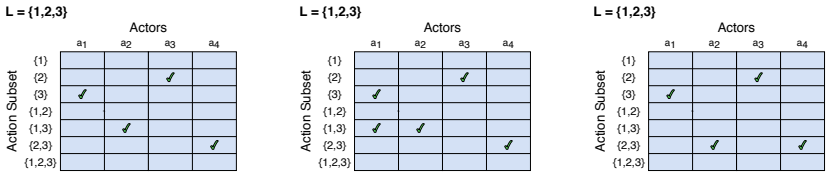
Since our goal is the assignment of a subset of actions ω to each actor, we do not need to compute the full probability (2). Instead of using the power set of all actions, we build the power set only for the actions that are provided as weak labels for each training video clip. This means that the power set will differ for each video clip. For the example shown in Figure 1, we build the power set Ω for the actions *Stand*, *Listen*, *Talk*, and *Watch*. In this example, $|\Omega| = 16$. We exclude $\varnothing$ since in the used dataset each actor is annotated with at least one action. Furthermore, we multiply p_ω with the confidence d of the person detector. The scoring function $p_{\omega,i}$ that we use for the assignment of a subset $\omega \in \Omega \setminus \varnothing$ to a detected actor a_i is therefore given by

$$p_{\omega,i} = \frac{\exp\left(\sum_{c\in\omega} s_{c,i}\right) d_i}{\sum_{\omega'\in\Omega\setminus\varnothing} \exp\left(\sum_{c\in\omega'} s_{c,i}\right)} \tag{3}$$

where $s_{c,i}$ is the predicted logit for action c and person a_i. Taking the detection confidence d_i of person a_i into account is necessary to reduce the impact of false positives that usually have a low detection confidence.

5.2 Actor-Action Association

While the scoring function (3) indicates how likely a given subset of actions $\omega \in \Omega \setminus \varnothing$ fits to an actor a_i, it does not take all information that is available for each video clip into account. For instance, we know that each annotated action is



a) Valid assignment. b) Invalid assignment. c) Invalid assignment.

Fig. 3. For the annotated actions $L = \{1, 2, 3\}$ and the actors $A = \{a_1, a_2, a_3, a_4\}$, the figures demonstrate various actor-action assignments. While the assignment a) satisfies all constraints, b) violates (5) since two subsets are assigned to actor a_1 and c) violates (6) since the action 1 is not part of any assigned subset.

performed by at least one actor. In order to exploit this additional knowledge, we find the optimal assignment of action subsets to actors based on the constraints that each actor performs at least one action and that each action c occurs at least once, i.e., $c \in \bigcup_i \omega_i$. Since we build the power set only from the actions that occur in a video clip, which we denote by L, the power set $\Omega(L)$ varies for each training video clip.

The association of subsets $\omega \in \Omega(L) \setminus \varnothing$ to actors $A = \{a_1, a_2, \ldots, a_n\}$ can be formulated as a binary linear program where the binary variables $x_{\omega,i}$ are one if the subset ω is assigned to actor a_i and it is zero otherwise. The optimal assignment is defined by the assignment with the highest score (4). While the first constraint (5) enforces that exactly one subset ω is assigned to each actor a_i, the second constraint (6) enforces that $c \in \bigcup_{\omega : x_{\omega,i}=1} \omega$ for all $c \in L$, where $\{\omega : x_{\omega,i} = 1\}$ is the set of all subsets that have been assigned. Note that (6) rephrases this constraint such that it can be used for optimization where the indicator function $\mathbb{1}_\omega(c)$ is one if $c \in \omega$ and it is zero otherwise. The left hand side of the inequality therefore counts the number of assigned subsets that contain the action class c. Since this number must be larger than zero, it ensures that each action $c \in L$ is assigned to at least one actor. The complete binary linear program is thus given by:

$$\underset{x_{\omega,i}}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{\omega \in \Omega(L) \setminus \varnothing} p_{\omega,i} x_{\omega,i} \tag{4}$$

$$\text{subject to} \sum_{\omega \in \Omega(L) \setminus \varnothing} x_{\omega,i} = 1 \qquad \forall i = 1, ..., n \tag{5}$$

$$\sum_{i=1}^{n} \sum_{\omega \in \Omega(L) \setminus \varnothing} \mathbb{1}_\omega(c) x_{\omega,i} \geq 1 \qquad \forall c \in L \tag{6}$$

$$x_{\omega,i} \in \{0, 1\} \qquad \forall \omega \in \Omega(L) \setminus \varnothing;\ \forall i = 1, ..., n.$$

Figure 3 illustrates the constraints.

5.3 Training

We train first the network using the MIML loss (1) to obtain initial estimates of the logits $s_{c,i}$. We then assign subsets of actions to the detected persons using the scoring function (3). Finally, we train our network using the loss

$$\mathcal{L} = \mathcal{L}_{MIML} + \alpha \sum_{i=1}^{n_t} \mathcal{L}\left(\hat{Y}_{\omega_i^t}, f(a_i^t)\right) \tag{7}$$

where ω_i^t denotes the action subset that has been assigned to actor a_i^t in frame t and $\hat{Y}_{\omega_i^t}$ is a vector with $\hat{Y}_{\omega_i^t}(c) = 1$ if $c \in \omega_i^t$ and $\hat{Y}_{\omega_i^t}(c) = 0$ otherwise. $\mathcal{L}$ is the binary cross entropy. Since $\mathcal{L}_{MIML}$ is computed once per frame but $\mathcal{L}(\hat{Y}_{\omega_i^t}, f(a_i^t))$ is computed for each detected person, we use $\alpha = 0.3$ to compensate for this difference.

6 Experiments

6.1 Dataset and Implementation Details

We use the AVA 2.2 dataset [19] for evaluation. The dataset contains 235 videos for training, 64 videos for validation, and 131 videos for testing. The dataset contains 60 action classes. The persons perform often multiple actions at the same time and the videos contain multiple persons. For each annotated person a bounding box is provided. An example is given in Figure 1. Only one frame per second is annotated. The accuracy is measured by mean average precision (mAP) over all actions with an IoU threshold for bounding boxes of 0.5 as described in [19]. In the weakly supervised setting, we use only the present actions for training, but not the bounding boxes.

To detect persons, we use Faster RCNN [41] with ResNext-101 [38] as backbone. The detector was pre-trained on ImageNet and fine-tuned on the COCO dataset. In our experiments, we report results for two 3D CNNs, namely I3D [39] and Slowfast [10]. I3D is pre-trained on Kinetics-400. For Slowfast, we use the ResNet-101 + NL (8×8) version that is pre-trained on Kinetics 600. The temporal scope was set to 64 frames with a stride of 2. For HGRNN we use a temporal window of 11 frames. For training, we use the SGD optimizer until the validation error saturated. The learning rate with linear warmup was set to 0.04 and 0.025 for I3D and Slowfast, respectively. The batch size was set to 16. We used cropping as data augmentation where we crop images of size 224×224 pixels from the frames that have 256×256 image resolution.[1]

6.2 Experimental Results

Comparison of MIML with proposed method. Table 1 shows the comparison of the proposed approach with the multi-instance and multi-label (MIML)

[1] Code: https://github.com/sovan-biswas/MultiLabelActorActionAssignment

Table 1. Comparison of MIML with proposed method. The proposed approach outperforms MIML in case of I3D and Slowfast.

Method	3D CNN	Val-mAP
MIML	I3D	14.1
MIML + HGRNN	I3D	15.2
Proposed Approach	I3D	17.3
MIML	Slowfast	21.8
MIML + HGRNN	Slowfast	23.1
Proposed Approach	Slowfast	25.1

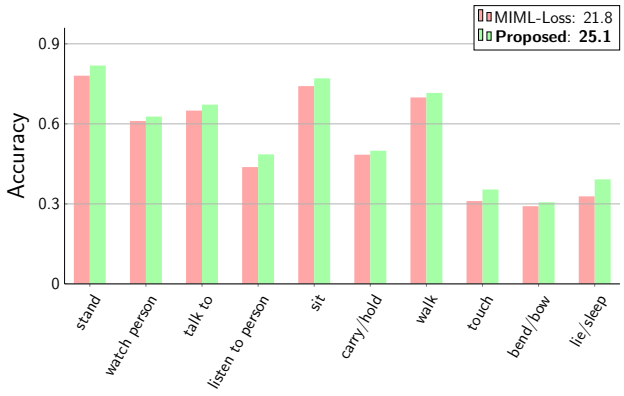


Fig. 4. Comparison of MIML with proposed method. The plot shows the per class mAP for the 10 most frequently occurring classes in the training set. The actions are sorted by the number of occurrences in an decreasing order from left to right. A plot with all 60 action classes is part of the supplementary material.

baseline on the validation set. When I3D is used as 3D CNN, the proposed approach improves the MIML baseline by +3.2%. When Slowfast is used, the accuracy of all methods is higher but the improvement of the proposed approach over the MIML approach remains nearly the same with +3.3%. We also report the result when HGRNN is trained only with the MIML loss. In this case, the actor-action association is not used and we denote this setting by MIML+HGRNN. While HGRNN improves the results since it models the spatio-temporal relations between persons better than a 3D CNN alone, the proposed actor-action assignment improves the mAP compared to MIML+HGRNN by +2.1% and +2.0% for I3D and Slowfast, respectively. Figure 4 shows the improvement of the proposed approach over the MIML baseline for the 10 action classes that occur most frequently in the training set. A few qualitative results are show in Figure 5.

Impact of actor-action association. In Table 1, we have observed that the actor-action association improves the accuracy. In Table 2, we analyze the impact

Table 2. Results of various actor-action assignment approaches using HGRNN over different 3D CNNs. The Frequent-5 column and the Least-10 column show the average mAP over the 5 most frequently and 10 least occurring classes in the training set.

Actor-action association	Backbone	Val-mAP	Frequent-5	Least-10
MIML+HGRNN	I3D	15.2	51.5	2.0
Proposed Approach w/o LP	I3D	16.4	52.8	2.1
Proposed Approach	I3D	**17.3**	**53.7**	**3.4**
MIML+HGRNN	Slowfast	23.1	65.7	7.3
Proposed Approach w/o LP	Slowfast	22.9	65.9	6.8
Proposed Approach	Slowfast	**25.1**	**67.5**	**7.6**

Table 3. Performance with ground-truth bounding boxes for evaluation. The results show the improvement in mAP on the validation set when ground-truth bounding boxes (GT bb) instead of detected bounding boxes (Detected bb) are used for evaluation. Furthermore, the results are reported when the model is trained with full supervision.

Method	3D CNN	Detected bb	GT bb
MIML	I3D	14.1	21.2
Proposed Approach	I3D	17.3	24.3
Full Supervision	I3D	20.7	25.4
MIML	Slowfast	21.8	30.6
Proposed Approach	Slowfast	25.1	32.3
Full Supervision	Slowfast	30.1	35.7

of the actor-action association more in detail. We use HGRNN using both I3D and Slowfast as 3D CNN backbone. In case of MIML+HGRNN, the actor-action association is not used. We also report the result if we perform the association directly by the confidences without solving a binary linear program. We denote this setting by Proposed Approach w/o LP. In this case, we associate an action to an actor if the class probability is greater than 0.5. For I3D, the association without LP improves the results mainly for the most frequent classes with almost no improvement on least frequent classes. For Slowfast, the performance even decreases in comparison to MIML+HGRNN without LP. Instead, solving the linear program results in better associations for both I3D and Slowfast.

Impact of the object detector. We use the Faster RCNN with ResNext [38] person detector which achieves 90.10% mAP for person detection on the AVA training set and 90.45% on the AVA validation set. Irrespective of the high detection performance, we analyze how much the accuracy improves if the detected bounding boxes are replaced with the ground-truth bounding boxes during evaluation. Note that the ground-truth bounding boxes are not used for training, but only for evaluation. The results are shown in Table 3. We observe that the performance improves by +7.0% and +7.2% mAP on the validation set for I3D and Slowfast, respectively. We also report the results if the approach is trained using full supervision. In this case, the network is trained on the ground-truth bound-

Table 4. Comparison to fully supervised approaches. We also report the result of our approach if it is trained with full supervision. Note that we do not use multi-scale and horizontal flipping augmentation as in Slowfast++.

Weakly Supervised Approaches		
Methods	Val-mAP	Test-mAP
MIML	21.8	-
Proposed Approach	**25.1**	**23.5**
Fully Supervised Approaches		
Methods	Val-mAP	Test-mAP
ARCN [5]	17.4	-
RAF [6]	20.4	-
HGRNN [7]	20.9	-
ATN [8]	25.0	24.9
LFB [9]	27.7	27.2
Slowfast [10]	29.0	-
Slowfast++ [10]	30.7	34.3
Proposed Approach	30.1	-

ing boxes and the ground-truth action labels per bounding box. Compared to the fully supervised approach, our weakly supervised approach achieves around 83% of the mAP for both 3D CNNs (17.3% vs. 20.7% for I3D and 25.1% vs. 30.1% for Slowfast) if detected bounding boxes are used for evaluation. The gap gets even smaller when ground-truth bounding boxes are used for evaluation. In this case, the relative performance is 95.7% for I3D and 90.5% for Slowfast. This demonstrates that the proposed approach learns the actions very well despite of the weak supervision.

Comparison to fully supervised methods. Since this is the first approach that addresses weakly supervised learning for multi-label and multi-person action detection, we cannot compare to other weakly supervised approaches. However, we compare our approach with the state-of-the-art for fully supervised action detection in Table 4. Our approach is competitive to fully supervised approaches [5–8]. When we train our approach with full supervision, we improve over SlowFast [10] by +1.1% mAP on the validation set. While the Slowfast++ network performs slightly better, it uses additional data augmentation and a different network configuration. We expect that these changes would improve our approach as well.

7 Conclusion

In this paper, we introduced the challenging task of weakly supervised multi-label spatio-temporal action detection with multiple actors. We first introduced a baseline based on multi-instance and multi-label learning. We furthermore presented a novel approach where the multi-label problem is represented by

the power set of the action classes. In this context, we assign an element of the power set to each detected person using linear programming. We evaluated our approach on the challenging AVA dataset where the proposed method outperforms the MIML approach. Despite of the weak supervision, the proposed approach is competitive to fully supervised approaches.

Acknowledgment

The work has been financially supported by the ERC Starting Grant ARCA (677650).

References

1. Gkioxari, G., Malik, J.: Finding action tubes. In: CVPR. (2015) 759–768
2. Hou, R., Chen, C., Shah, M.: Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos. In: ICCV. (2017) 5822–5831
3. Kalogeiton, V., Weinzaepfel, P., Ferrari, V., Schmid, C.: Action Tubelet Detector for spatio-temporal action localization. In: ICCV. (2017) 4415–4423
4. Singh, G., Saha, S., Sapienza, M., Torr, P., Cuzzolin, F.: Online Real-Time multiple spatiotemporal action localisation and prediction. In: ICCV. (2017) 3657–3666
5. Sun, C., Shrivastava, A., Vondrick, C., Murphy, K., Sukthankar, R., Schmid, C.: Actor-Centric Relation Network. In: ECCV. (2018) 318–334
6. Sun, C., Shrivastava, A., Vondrick, C., Sukthankar, R., Murphy, K., Schmid, C.: Relational action forecasting. In: CVPR. (2019)
7. Biswas, S., Souri, Y., Gall, J.: Hierarchical Graph-RNNs for Action Detection of Multiple Activities. In: ICIP. (2019)
8. Girdhar, R., Carreira, J., Doersch, C., Zisserman, A.: Video Action Transformer Network. In: CVPR. (2019) 244–253
9. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: CVPR. (2019) 284–293
10. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast Networks for Video Recognition. In: ICCV. (2019) 6202–6211
11. Mettes, P., Snoek, C.G., Chang, S.F.: Localizing actions from video labels and pseudo-annotations. In: BMVC. (2017)
12. Soomro, K., Shah, M.: Unsupervised action discovery and localization in videos. In: ICCV. (2017) 696–705
13. Chéron, G., Alayrac, J.B., Laptev, I., Schmid, C.: A flexible model for training action localization with varying levels of supervision. In: NIPS. (2018) 942–953
14. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. (2012)
15. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV. (2013) 3192–3199
16. Zhou, Z.H., Zhang, M.L.: Multi-instance multi-label learning with application to scene classification. In: NIPS. (2006) 1609–1616
17. Zhou, Z.H., Zhang, M.L., Huang, S.J., Li, Y.F.: Multi-instance multi-label learning. Artificial Intelligence **176** (2012) 2291–2320
18. Yang, H., Tianyi Zhou, J., Cai, J., Soon Ong, Y.: MIML-FCN+: Multi-instance multi-label learning via fully convolutional networks with privileged information. In: CVPR. (2017) 1577–1585

19. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: AVA: A Video dataset of Spatio-Temporally Localized Atomic Visual Actions. In: CVPR. (2018) 6047–6056
20. Song, L., Zhang, S., Yu, G., Sun, H.: TACNet: Transition-Aware Context Network for Spatio-Temporal Action Detection. In: CVPR. (2019) 11987–11995
21. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: CVPR. (2020)
22. Ji, J., Krishna, R., Fei-Fei, L., Niebles, J.C.: Action genome: Actions as compositions of spatio-temporal scene graphs. In: CVPR. (2020)
23. Weinzaepfel, P., Martin, X., Schmid, C.: Towards weakly-supervised action localization. arXiv preprint arXiv:1605.05197 (2016)
24. Siva, P., Xiang, T.: Weakly Supervised Action Detection. In: BMVC. (2011) 6
25. Mettes, P., Snoek, C.G.: Spatio-temporal instance learning: Action tubes from class supervision. arXiv preprint arXiv:1807.02800 (2018)
26. Ghadiyaram, D., Tran, D., Mahajan, D.: Large-scale weakly-supervised pre-training for video action recognition. In: CVPR. (2019)
27. Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Finding actors and actions in movies. In: ICCV. (2013) 2280–2287
28. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008) 1–8
29. Kuehne, H., Richard, A., Gall, J.: A hybrid RNN-HMM approach for weakly supervised temporal action segmentation. PAMI (2018)
30. Richard, A., Kuehne, H., Gall, J.: Action sets: Weakly supervised action segmentation without ordering constraints. In: CVPR. (2018) 5987–5996
31. Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A., Murphy, K., Fei-Fei, L.: Detecting events and key actors in multi-person videos. In: CVPR. (2016) 3043–3053
32. Li, J., Liu, J., Yongkang, W., Nishimura, S., Kankanhalli, M.: Weakly-supervised multi-person action recognition in 360° videos. In: WACV. (2020)
33. Nguyen, C.T., Zhan, D.C., Zhou, Z.H.: Multi-modal image annotation with multi-instance multi-label lda. In: IJCAI. (2013)
34. Nguyen, N.: A new SVM approach to multi-instance multi-label learning. In: ICDM. (2010) 384–392
35. Briggs, F., Fern, X.Z., Raich, R.: Rank-loss support instance machines for MIML instance annotation. In: SIGKDD. (2012) 534–542
36. Zha, Z.J., Hua, X.S., Mei, T., Wang, J., Qi, G.J., Wang, Z.: Joint multi-label multi-instance learning for image classification. In: CVPR. (2008) 1–8
37. Zhang, X.Y., Shi, H., Li, C., Li, P.: Multi-instance multi-label action recognition and localization based on spatio-temporal pre-trimming for untrimmed videos. In: AAAI. (2020) 12886–12893
38. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR. (2017)
39. Carreira, J., Zisserman, A.: Quo Vadis, Action recognition? A new model and the kinetics dataset. In: CVPR. (2017) 4724–4733
40. Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., Adam, H.: Large-scale object classification using label relation graphs. In: ECCV. (2014) 48–64
41. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: NIPS. (2015) 91–99

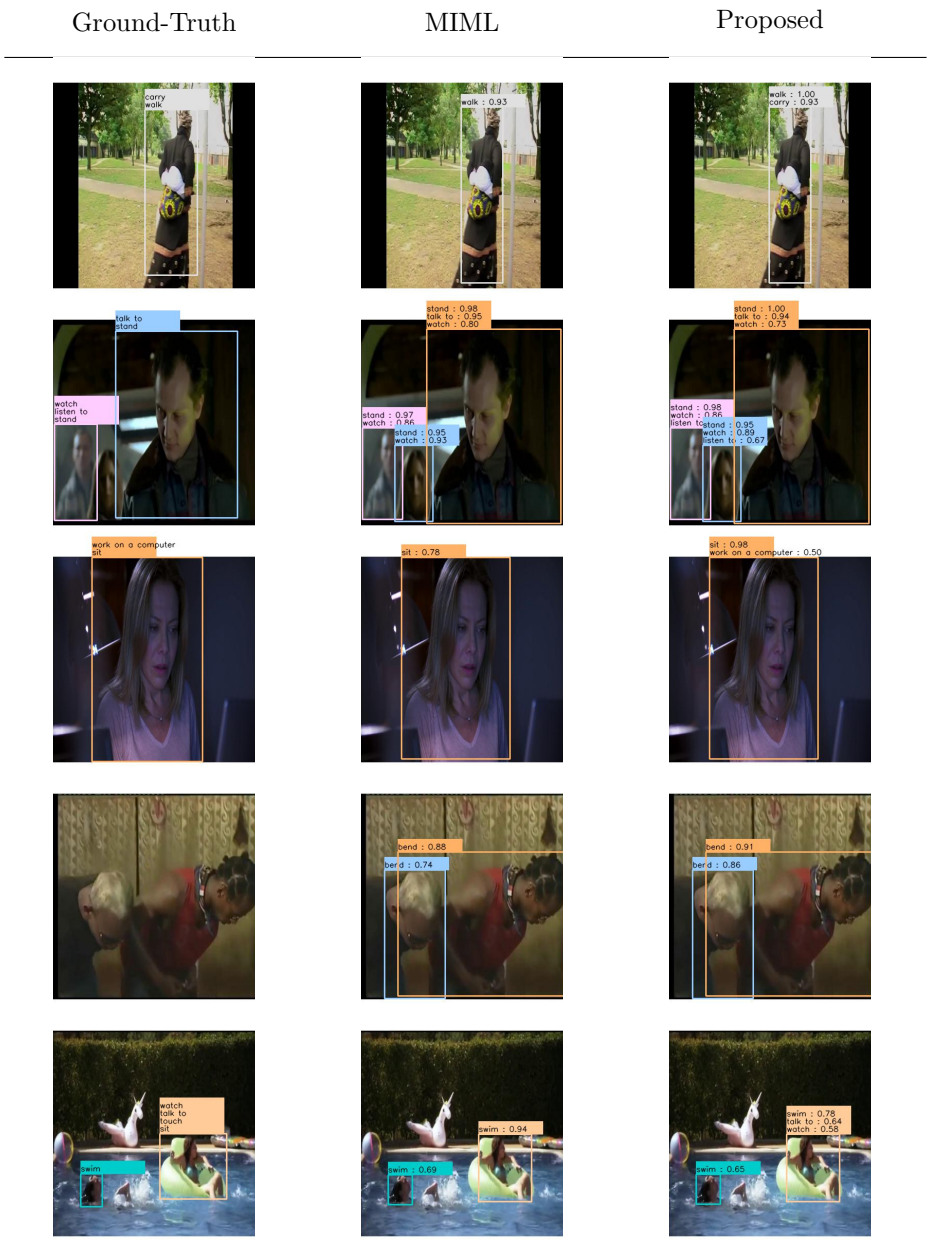


Fig. 5. Qualitative results. The left column shows the ground-truth annotations. The middle column shows the results of the MIML baseline. The right column shows the results of the proposed method. The colors distinguish only different persons, but they are otherwise irrelevant. The predicted action classes with confidence scores are on top of the estimated bounding boxes. The proposed approach recognizes more action classes per bounding box correctly compared to MIML. Both methods also detect genuine actions that are not annotated in the dataset as seen from the missing persons in the second and fourth row. The bias of the proposed method towards the background is visible in last row, where the "swim" action is associated to both persons. Best viewed using the zoom function of the PDF viewer.