Multi-Instance Multi-Label Action Recognition and Localization Based on Spatio-Temporal Pre-Trimming for Untrimmed Videos

Xiao-Yu Zhang,[1*] **Haichao Shi,**[1,2*] **Changsheng Li,**[3†] **Peng Li**[4†]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]School of Computer Science and Technology, Beijing Institute of Technology
[4]China University of Petroleum (East China)
{zhangxiaoyu, shihaichao}@iie.ac.cn, changsheng_li_507@hotmail.com, lipeng@upc.edu.cn

Abstract

Weakly supervised action recognition and localization for untrimmed videos is a challenging problem with extensive applications. The overwhelming irrelevant background contents in untrimmed videos severely hamper effective identification of actions of interest. In this paper, we propose a novel multi-instance multi-label modeling network based on spatio-temporal pre-trimming to recognize actions and locate corresponding frames in untrimmed videos. Motivated by the fact that person is the key factor in a human action, we spatially and temporally segment each untrimmed video into person-centric clips with pose estimation and tracking techniques. Given the bag-of-instances structure associated with video-level labels, action recognition is naturally formulated as a multi-instance multi-label learning problem. The network is optimized iteratively with selective coarse-to-fine pre-trimming based on instance-label activation. After convergence, temporal localization is further achieved with local-global temporal class activation map. Extensive experiments are conducted on two benchmark datasets, i.e. THUMOS14 and ActivityNet1.3, and experimental results clearly corroborate the efficacy of our method when compared with the state-of-the-arts.

Introduction

Action recognition and localization is a heated research topic with intensive attention for high-level video understanding. Compared with trimmed videos which are temporally aligned with specific actions of interest, untrimmed videos are typically flooded with irrelevant background contents, and thus far more challenging to analyze. In the meantime, frame-level full supervision is accompanied with prohibitive labeling effort, and apparently unfeasible for large scale video sets. As a result, we focus on the more practical weak supervision scenario, and aims to identify actions and the corresponding temporal intervals in untrimmed videos with video-level annotation.

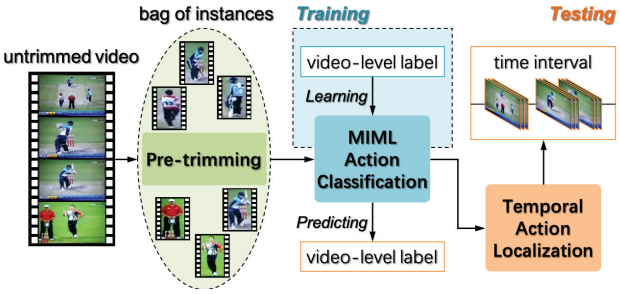


Figure 1: The workflow overview of PreTrimNet.

The overwhelming action-irrelevant contents in untrimmed videos are formidable obstacles to effective action identification. Directly modeling the untrimmed video as a whole can be quite misleading, and will inevitably result in deteriorated performance. As we know, trimmed videos are noise-free high-quality data to develop accurate action recognizers. Likewise, it is more reasonable to "denoise" the untrimmed video by eliminating the action-irrelevant contents and retaining segmented clips that are dominated by potential actions. Traditional clip generation methods (such as those with sliding windows, pre-defined durations, flexible boundaries, etc.) are mostly based on low-level features and only confined to temporal segmentation, leaving the spatial contents unmodified. This paper aims to explore spatio-temporal segmentation of untrimmed videos, with more reliable evidence. As we know, a human action is closely related to the person that is engaged in the movement. By spatially and temporally locating the key persons, untrimmed videos can be segmented into person-centric clips accordingly, based on which model learning can be implemented in a more reliable way. Generally speaking, an untrimmed video corresponds to multiple person-centric clips, each of which focuses on a certain person performing a specific action. Taking the untrimmed video as a bag and the multiple clips as instances in the bag, action recognition can be conveniently formulated as a multi-instance multi-label learning problem, which is inherently proficient in discovering instance-label relation.

Motivated by the above considerations, in this paper,

*These authors contributed equally to this study and share the first authorship.
†Corresponding authors.

we propose a weakly supervised action recognition and localization method based on multi-instance multi-label modeling with spatio-temporal pre-trimming, referred to as PreTrimNet for short, which is illustrated in Figure 1. Untrimmed videos are represented as bags of clips corresponding to specific persons via pre-trimming. In the training stage, the network learns from the given video-level labels to construct action recognition model. In the testing stage, the optimized network aims to estimate both the actions of interest and their temporal intervals with unlabeled input. The main contributions of our work are summarized as follows.

- To the best of our knowledge, PreTrimNet is the first to present spatio-temporal pre-trimming based on pose estimation and tracking to generate person-centric clips from untrimmed videos for action recognition and localization with weak supervision, which updates in a coarse-to-fine fashion with the iteration of learning model.
- Under the multi-instance multi-label learning mechanism, PreTrimNet is carefully designed to discover the underlying relevance between input patterns and semantic labels via integration of three-stream features, based on which temporal localization is further achieved from a local-global perspective.
- Extensive experiments on two challenging untrimmed video datasets, i.e. THUMOS14 (Jiang et al. 2014) and ActivityNet1.3 (Heilbron et al. 2015), show promising results of PreTrimNet over the existing state-of-the-art competitors. Note that we focus on human actions in this paper. However, PreTrimNet can be easily generalized to non-human actions.

Related Work

Action Recognition and Localization

Action recognition is conventionally formulated as a classification problem which aims to determine the categories of human actions in a video. Before the prevalence of deep learning, hand-crafted features, such as the improved dense trajectories (iDT) (Wang and Schmid 2013), have obtained outstanding performance on benchmark datasets. During the last few years, deep architectures have been successfully applied to video-based action analysis. Two-stream (Simonyan and Zisserman 2014) and C3D (Tran et al. 2015) networks are recent mainstreams to learn discriminative features. The inception 3D (I3D) (Carreira and Zisserman 2017) is exploited to use a two-stream network based on a 3D version of Inception network (Ioffe and Szegedy 2015).

Temporal action localization aims to identify the temporal intervals which contain target actions. S-CNN (Shou, Wang, and Chang 2016) utilizes a multi-stage CNN to learn robust feature representation. Boundary Sensitive Network (BSN) (Lin et al. 2018) generates temporal proposals by probability estimation. Recently, action localization in weakly supervised fashion has gained increasing attention. UntrimmedNet (Wang et al. 2017) is trained end-to-end for single-label action recogniton and localization. Hide-and-seek (Singh and Lee 2017) enforces the model to see different parts of the image and focuses on multiple relevant parts of the object by randomly masking different regions of training images. STPN (Nguyen et al. 2018) adopts an attention module to identify a sparse subset of key segments associated with actions. AutoLoc (Shou et al. 2018) directly predicts the temporal boundary of each action instance with an outer-inner-contrastive loss. W-TALC (Paul, Roy, and Roy-Chowdhury 2018) learns the specific network weights by optimizing co-activity similarity and multiple instance learning loss. TSRNet (Zhang et al. 2019) integrated transfer learning by leveraging knowledge from external trimmed videos.

Pose Estimation and Tracking

Pose estimation tasks can be divided into single person and multiple person scenarios. Single person pose estimation attempts to estimate the pose of a single person. Conventional methods adopt pictorial structure models, such as tree models (Sapp, Toshev, and Taskar 2010; Wang and Li 2013) and random forest models (Sun, Kohli, and Shotton 2012). Representative deep learning based single person pose estimation methods include DeepPose (Toshev and Szegedy 2014) and CNN based methods (Belagiannis and Zisserman 2017). Multi-person pose estimation is a lot more challenging. Part-based framework investigates local regions to detect human body parts. Chen et al. (2015) propose to use graphical model to model humans as flexible compositions of body parts. Pishchulin et al. (2016) propose DeepCut to first detect all body parts and then label and assemble these parts via integral linear programming. Two-step framework has achieved further improvement. Fang et al. (2017) propose to use a CNN based SPPE method to estimate poses namely *AlphaPose*. Pishchulin et al. (2012) use conventional pictorial structure models for pose estimation. Insafutdinov et al. (2016) propose a two-step pipeline which uses the Faster R-CNN as their human detector and a unary DeeperCut as their pose estimator.

Multi-instance Multi-label learning

In many real-world applications, the objects of interest are typically with multiple labels, and can be represented as a bag of multiple instances. Many successful multi-instance multi-label (MIML) learning algorithms have been proposed. Huang et al. (2014) propose a fast MIML algorithm by exploiting label relations with shared space and discovering sub-concepts for complicated labels. Pham et al. (2015) use a discriminative probabilistic model to discover novel class instances in a MIML setting. Recently, deep learning based MIML algorithms have achieved great success in various tasks, such as multi-label image classification (Wang et al. 2016a), object detection (Ge, Yang, and Yu 2018). Specifically, Simonyan et al. (2015) utilize MIML to classify images based on deep covolutional neural networks. Kiros et al. (2015) use a pre-trained skip-thought model in an off-the-shelf encoder to produce high quality sentence representations. Feng et al. (2017) propose DeepMIML network which exploits deep neural network formation for MIML, and exhibits robust instance-label relation discovery ability.

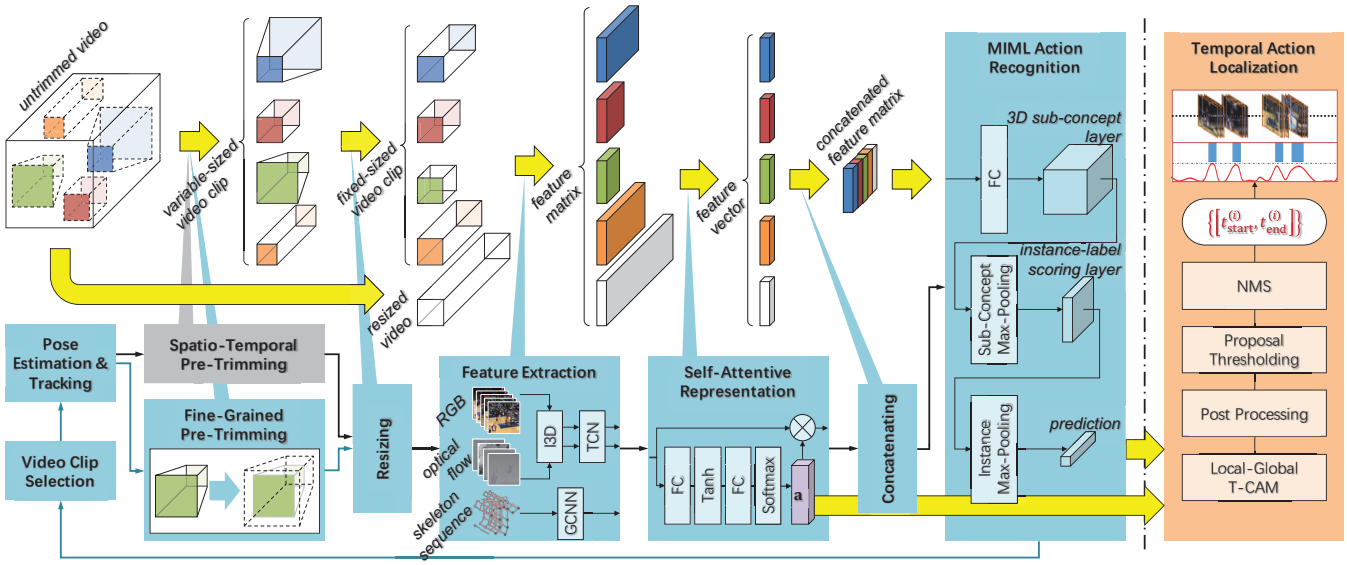


Figure 2: The detailed framework of PreTrimNet (better viewed in color).

Proposed Method

In this section, we present the proposed PreTrimNet in detail, whose methodological framework is illustrated in Figure 2. The model starts with spatio-temporal pre-trimming based on pose estimation and tracking. Accordingly, an untrimmed video is segmented into multiple clips corresponding to specific persons. After resizing, the fixed-sized clips are fed into feature extraction and self-attentive representation modules to obtain compact feature vectors, which are subsequently concatenated into a feature matrix. Multi-instance multi-label learning is used to formulate and train the action recognition model given video-level labels. Based on the instance-label activation, coarse-to-fine pre-trimming is implemented and the model is updated in an iterative way. Finally, after convergence, PreTrimNet is capable of predicting actions in unlabeled videos, as well as identifying the corresponding time intervals via a two-stage temporal class activation map.

Spatio-Temporal Pre-trimming

In order to effectively identify the actions of interest in an untrimmed video, it is essential to eliminate the irrelevant backgrounds and only retain the informative parts. Temporally, the untrimmed video should be segmented according to time intervals corresponding to potential actions, which is similar to the process of acquiring trimmed videos. Apart from that, we further trim the video spatially so as to highlight region of the specific action in each frame.

As discussed above, pose is a strong clue to identify human actions. Based on multi-person pose estimation and tracking techniques, we devise the spatio-temporal pre-trimming module to generate person-centric clips of an untrimmed video. The duration of each pose sequence associated with a person can be leveraged to determine the clip's time interval. Within each frame of the clip, the pose bounding box in which the person dominates the visual content is extracted, with the rest removed. Due to camera zooming, we arrive at a set of variable-sized clips, among which the less important persons whose poses last less than a preset period of time are filtered out. For the sake of computational efficiency, we also delete clips of static persons, such as audiences, which contain little motion information. Finally, all the clips are uniformly resized so that they can conveniently go through unified posterior processing. Let $\mathcal{V} = \{v_i|_{i=1}^{M}\}$ denote the set of clips extract from a given untrimmed video $\mathcal{V}$, where M is the number of clips in $\mathcal{V}$. With spatio-temporal pre-trimming, the untrimmed video is divided into high-quality clips which are closely related to actions of interest.

Three-Stream Feature Extraction

To fully capture both visual and motional information from the pre-trimmed video clips, it has been a standard practice to utilize the two-stream architecture to derive spatial and temporal descriptions. As the name suggests, video features are extracted via two separately trained networks corresponding to RGB and optical flow, respectively. As for the feature extractor, we employ the I3D network pre-trained on Kinetics for both streams, and receive frame-level clip features. To be specific, the input are non-overlapping frame chunks, and the output is passed through a 3D average pooling layer to obtain d-dimensional features. The I3D block captures the short-range temporal patterns within the vicinity of adjacent frames. To further depict the long-range dependencies, the two-stream features are fed subsequently into the TCN (Temporal Convolutional Networks), which is comprised of a hierarchy of temporal convolutions.

Besides the traditional two-stream features, we further make full use of the pose sequences derived from pose estimation and tracking. Since each clip is pre-trimmed to con-

tain only one person, single-person skeleton sequence representation learning model, such as GCNN, is leveraged to extract pose feature. Compared with RGB and optical flow, pose conveys higher-level information of the video contents, and thus offers a new perspective for video description and understanding.

Formally, given a clip $v_i \in \mathcal{V}$, the feature from one of the three streams is denoted as a d_*-by-t_* dimensional matrix $\mathbf{X}_i^*$, where $* \in \{\mathrm{RGB}, \mathrm{FLOW}, \mathrm{POSE}\}$ indicates the specific feature stream, d_* is the feature dimension, and t_* is dependent on the feature extraction procedure and proportional to the number of frames in v_i.

Compact Representation via Self-Attention

The variable lengths of extracted clips bring about variable-sized feature matrices, which are extremely inconvenient to process. In order to get fixed-sized compact representations for the clips, we leverage the self-attention mechanism to integrate the frame-level descriptions. The self-attention block consists of four parts, i.e. two fully connected (FC) layers, a tanh activation layer between the two FC layers, and a softmax activation layer to ensure each set of generated weights sum up to 1. By modeling the global dependency of the frame sequence in a clip v_i regardless of the inter-frame distances, a linear combination of column vectors in the feature matrix $\mathbf{X}_i^*$ is learned as $\mathbf{f}_i^* = \mathbf{X}_i^* \mathbf{a}_i^*$, where $\mathbf{f}_i^* \in \mathbb{R}^{d_* \times 1}$ is the compact vector representation for clip v_i, and $\mathbf{a}_i^* \in \mathbb{R}^{t_* \times 1}$ is the corresponding attention weight vector calculated as:

$$\mathbf{a}_i^* = (\mathrm{softmax}(\mathbf{w}_2 \mathrm{tanh}(\mathbf{W}_1 \mathbf{X}_i^*)))^\top \tag{1}$$

where $\mathbf{W}_1 \in \mathbb{R}^{b \times d}$ and $\mathbf{w}_2 \in \mathbb{R}^{1 \times b}$ are intermediate parameters to be learned, and b is a hyperparameter set empirically.

To ensure robustness of the self-attentive representations, we impose adjacency smoothness constraint on attention weights to follow the vicinity resemblance property of video frames, which takes on the following form:

$$\mathcal{L}_{\mathrm{smoothness}} = \sum_{v_i \in \mathcal{V}} \sum_{j=1}^{t-1} |a_{i,j}^* - a_{i,j+1}^*|^2 \tag{2}$$

Specifically, when the input is the entire untrimmed video rather pre-trimmed clips, we should further impose sparsity constraint to enforce the insignificant weights to be 0 so that the negative impact of irrelevant background contents can be naturally eliminated. As will be discussed in the next subsection, we use the untrimmed video itself as the $(M+1)$-th clip, and thus the additional sparsity constraint is designed as:

$$\mathcal{L}_{\mathrm{sparsity}} = ||\mathbf{a}_{M+1}^*||_{2,1} \tag{3}$$

After self-attentive representation learning, three-stream feature vectors can be stacked together to generate a comprehensive feature vector $\mathbf{h}_i = [\mathbf{f}_i^{\mathrm{RGB}}; \mathbf{f}_i^{\mathrm{FLOW}}; \mathbf{f}_i^{\mathrm{POSE}}] \in \mathbb{R}^{d \times 1}$, where $d = d_{\mathrm{RGB}} + d_{\mathrm{FLOW}} + d_{\mathrm{POSE}}$.

Multi-instance Multi-label Action Recognition

So far, for each untrimmed video, we have got a set of person-centric clips as well as a set of class labels on the video level. Taking the untrimmed video as a bag and the spatio-temporal pre-trimmed clips as instances in the bag, action recognition can be naturally formulated as a multi-instance multi-label (MIML) learning problem. We design our MIML action recognition module based on the DeepMIML network, which has proven effective for tasks in various domains.

By concatenating the fixed-sized self-attentive representations of clips in an untrimmed video, we obtain a matrix description for the bag of instances. Note that, except for the pre-trimmed clips, we further incorporate the original untrimmed video as a special instance for two considerations.

- On one hand, by definition, multi-instance learning requires that a bag is positive for a label if at least one instance in it is positive. However, if the pre-trimming is unreliable, it is possible that none of the segmented clips contains the video-level actions. In this case, the additional instance corresponding to the untrimmed video will guarantee the legitimacy of the MIML algorithm.
- On the other hand, since the untrimmed video contains all the original frames, its class-agnostic attention weights as well as class-specific activation scores can serve as a supplementary global indicator to the local information from clips. With the help of the unified global evaluation on the additional instance, pre-trimming can be further refined and frame-level prediction can be further augmented.

Therefore, the input corresponding to untrimmed video $\mathcal{V}$ is a d-by-$(M+1)$ dimensional feature matrix $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_M, \mathbf{h}_{M+1}]$. Here we assume each untrimmed video has equal number of clips; for those with different number of clips, zero paddings may be applied.

Matrix $\mathbf{H}$ is fed into a fully connected 3D sub-concept layer of size $S \times C \times M$, where S and C are the number of sub-concepts and action categories, respectively. The activation of the (s, c, i)-th class represents the matching score of the s-th sub-concept of the c-th class for the i-th clip v_i in untrimmed video $\mathcal{V}$. Formally, the (s, c, i)-th node has the following form of activation:

$$\sigma_{s,c,i} = f(\mathbf{w}_{s,c} \mathbf{h}_i + b_{s,c}) \tag{4}$$

where $f(\cdot)$ is the activation function, which is ReLU (Rectified Linear Unit) in this paper.

The 3D sub-concept layer is followed by two pooling operations. Concretely, we first conduct max-pooling along the sub-concept dimension, and arrive at a 2-dimensional instance-label scoring layer of size $C \times M$. The activation of the (c, i)-th node represents the matching score for clip $\mathbf{h}_i$ on label c, based on which we can conveniently discover the relation between input low-level patterns and output high-level semantics.

After that, a second max-pooling along the instance dimension is implemented and produce a 1-dimensional prediction layer of size $C \times 1$. This can be interpreted as the matching scores for labels on video level.

Finally, given a training set $\mathcal{T}$ of untrimmed videos with video-level labels, classification model can be trained by minimizing the classification loss $\mathcal{L}_{\mathrm{CLS}}$, which is computed

with the standard multi-label cross-entropy loss. Based on (2), (3) and $\mathcal{L}_{\text{class}}$, we arrive the overall loss as follows:

$$\mathcal{L} = \sum_{\mathcal{V}\in\mathcal{T}} (\mathcal{L}_{\text{CLS}} + \varepsilon(\mathcal{L}_{\text{smoothness}} + \mathcal{L}_{\text{sparsity}})) \quad (5)$$

where ε is the trade-off parameter.

Iterative Coarse-to-Fine Pre-trimming

Using the MIML action recognition predictions as instructive hints, we can further refine the pre-trimming of some clips selectively. To be specific, based on activations in the instance-label scoring layer, we can locate the key clips that trigger the video-level label. The pre-trimming quality of these key clips will mostly account for the model's overall performance. As a result, we focus on the key clips, and find their outer boundary based on the global class-specific activation scores of $\mathbf{h}_{M+1}$. Then pose estimation and tracking are implemented on the expanded clips instead of the untrimmed videos to obtain finer-grained pose sequences which consequently bring about finer-grained spatio-temporal pre-trimming. Afterwards, for the sake of efficiency, we will only carry out pre-trimming when new key clips emerge.

Temporal Action Localization

Based on action recognition, the video-level labels are revealed. To further identify the time intervals in untrimmed videos corresponding to the actions of interest, the frame-level class-specific relevance should be investigated. As introduced in the MIML action recognition module, the predictive label for clip v_i is determined by $\max_s \sigma_{s,c,i}$.

$$\begin{aligned}\max_s \sigma_{s,c,i} &\propto \max_s \mathbf{w}_{s,c}\mathbf{h}_i \\ &= \max_s \sum_{*\in\{\text{RGB,FLOW,POSE}\}} \mathbf{w}^*_{s,c}\mathbf{f}^*_i\end{aligned} \quad (6)$$

where $\mathbf{w}^*_{s,c}$ is the weight vector in $\mathbf{w}_{s,c}$ corresponding to feature stream $*$. We define the class-specific score $g_{c,i}$:

$$\begin{aligned}g^*_{c,i} &= \mathbf{w}^*_{\tilde{s},c}\mathbf{f}^*_i \\ &= \mathbf{w}^*_{\tilde{s},c}\mathbf{X}^*_i\mathbf{a}^*_i \\ &= [\mathbf{w}^*_{\tilde{s},c}\mathbf{x}^*_{i,1}, ..., \mathbf{w}^*_{\tilde{s},c}\mathbf{x}^*_{i,t}]\mathbf{a}^*_i\end{aligned} \quad (7)$$

where $\tilde{s} = \operatorname{argmax}_s \sigma_{s,c,i}$, and $\mathbf{x}^*_{i,k}$ is feature of the k-th frame ($1 \le k \le t$). Based on T-CAM (Temporal Class Activation Map), the relevance between the k-th frame of the i-th clip with the c-th class can be evaluated with $p^*_{c,i,k} = \mathbf{w}^*_{\tilde{s},c}\mathbf{x}^*_{i,k}$.

To be specific, we propose the local-global T-CAM for temporal localization. Given an untrimmed video $\mathcal{V} = \{v_i|_{i=1}^M\}$, since the clips segmented with pose information is most likely to contain actions of interest, we firstly conduct local-evaluation for the frame-class relevance within M clips. Then using the predictive score of the additional $(M+1)$-th instance, which corresponds to the entire video, global-evaluation is made to identify potentially missed actions from backgrounds, filter out false actions in the less important clips, and further refine the action boundaries. Finally, temporal localization is achieved via NMS (Non-Maximum Suppression).

Table 1: Classification accuracy (%) on the THUMOS14 dataset for action recognition.

Method	RGB	Optical Flow	Pose	Fusion
(Wang and Schmid 2013)	-	-	-	63.1
(Wang et al. 2016b)	-	-	-	78.5
(Wang et al. 2017)	-	-	-	82.2
(Zhang et al. 2019)	74.4	79.6	-	87.1
PreTrimNet	**81.1**	**83.7**	**76.2**	**89.2**

Table 2: Classification accuracy (%) on the ActivityNet1.3 dataset for action recognition.

Method	RGB	Optical Flow	Pose	Fusion
(Zhang et al. 2019)	79.7	84.3	-	91.2
PreTrimNet	**85.1**	**87.4**	**81.7**	**93.3**

Experiments

In this section, we report the experimental evaluation of the proposed PreTrimNet on action recognition and localization tasks, in comparison with other state-of-the-art methods based on both fully and weakly supervised learning.

Datasets

We evaluate different methods on two benchmark datasets, i.e. THUMOS14 and ActivityNet1.3. Both datasets contain large numbers of untrimmed videos attached with temporal annotations of actions. Note that, as a weakly supervised model, PreTrimNet only has access to the video-level labels in the training stage.

THUMOS14. The THUMOS14 dataset contains a validation set of 1,010 videos and a testing set of 1,574 videos, which fall into 101 action classes. Among these videos, we focus on the temporally annotated 20-class subset. We train the model on the validation set of 200 videos, and evaluate it with the testing set of 213 videos. THUMOS14 is challenging in that it contains videos with multiple actions.

ActivityNet1.3. The ActivityNet1.3 dataset is originally comprised of 200 activity classes, with 10,024 videos for training, 4,926 for validation, and 5,044 for testing. Since the ground-truth labels for the original testing set are withheld, we adopt the training set for model training and the validation set for testing. ActivityNet1.3 contains a large number of natural videos that involve various human activities.

Implementation and Evaluation Details

We utilize the two-stream I3D networks pre-trained on Kinetics dataset to extract the traditional two-stream features. For the RGB stream, we perform the center crop of size 224 $\times$ 224. For the optical flow stream, we apply the TV-$L1$ optical flow algorithm. The input to the I3D models are stacks

Table 3: Comparison of action localization results on THUMOS14. (Methods in the upper and lower parts are with full and weak supervision, respectively.)

Method	mAP@IoU (%)								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
(Shou, Wang, and Chang 2016)	47.7	43.5	36.3	28.7	19.0	10.3	5.3	-	-
(Yeung et al. 2016)	48.9	44.0	36.0	26.4	17.1	-	-	-	-
(Yuan et al. 2016)	51.4	42.6	33.6	26.1	18.8	-	-	-	-
(Xu and Das 2017)	54.5	51.5	44.8	35.6	28.9	-	-	-	-
(Zhao et al. 2017)	**66.0**	**59.4**	51.9	41.0	29.8	-	-	-	-
(Lin et al. 2018)	-	-	**53.5**	45.0	36.9	28.4	20.0	-	-
(Chao et al. 2018)	59.8	57.1	53.2	**48.5**	**42.8**	**33.8**	**20.8**	-	-
(Singh and Lee 2017)	36.4	27.8	19.5	12.7	6.8	-	-	-	-
(Wang et al. 2017)	44.4	37.7	28.2	21.1	13.7	-	-	-	-
(Nguyen et al. 2018)	45.3	38.8	31.1	23.5	16.2	9.8	5.1	2.0	0.3
(Nguyen et al. 2018)	52.0	44.7	35.5	25.8	16.9	9.9	4.3	1.2	0.1
(Shou et al. 2018)	-	-	35.8	29.0	21.2	13.4	5.8	-	-
(Paul, Roy, and Roy-Chowdhury 2018)	55.2	49.6	40.1	31.1	22.8	-	7.6	-	-
(Su, Zhao, and Lin 2018)	47.1	41.6	32.8	24.7	16.1	10.1	5.5	-	-
(Zhang et al. 2019)	55.9	46.9	38.3	28.1	18.6	11.0	5.59	2.19	0.29
PreTrimNet	**57.49**	**50.73**	**41.40**	**32.05**	**23.09**	**14.16**	**7.69**	**2.33**	**0.39**

Table 4: Comparison of action localization results on the ActivityNet1.3. (Methods in the upper and lower parts are with full and weak supervision, respectively.)

Method	mAP@IoU (%)			
	0.5	0.75	0.95	Average
(Xu and Das 2017)	26.8	-	-	-
(Heilbron et al. 2017)	40.0	17.9	4.7	21.7
(Shou et al. 2017)	45.3	26.0	0.2	23.8
(Zhao et al. 2017)	39.12	23.48	5.49	23.98
(Lin et al. 2018)	**52.50**	**33.53**	**8.85**	**33.72**
(Nguyen et al. 2018)	29.3	16.9	2.6	-
(Zhang et al. 2019)	33.1	18.7	3.32	21.78
PreTrimNet	**34.8**	**20.9**	**5.3**	**22.5**

of 16 (RGB or flow) frames sampled at 16 frames per second. Graph based method is used to extract pose features from skeleton sequences. Feature dimensions for RGB, optical flow, and pose are 1,024, 1,024, and 512, respectively. The model parameters are optimized with the mini-batch stochastic gradient descent with Adam optimizer. The learning rate is set to 0.0001 for the RGB and pose stream and decreases every 3,000 iterations by a factor of 10. For the optical flow stream, we set the learning rate to 0.0005, which is decreased every 5,000 iterations by a factor of 10. We also utilize the dropout operations with ratios to 0.5 and common augmentation techniques including horizontal flipping, cropping augmentation, et al. Our algorithm is implemented in PyTorch.

We follow the standard evaluation metric, which is based on the values of mean average precision (mAP) under different levels of intersection over union (IoU) thresholds.

Table 5: Ablation study results on THUMOS14 of temporal action localization.

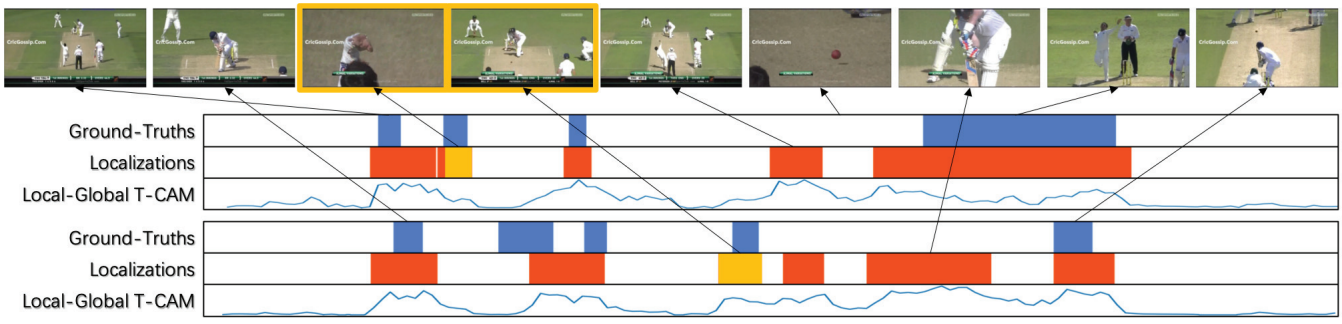
Method	Avg(0.1:0.5)
PreTrimNet w/o MIML, Pre-trimming	29.26
PreTrimNet w/o Pre-trimming	32.40
PreTrimNet w/o MIML	36.31
PreTrimNet	**40.95**

Results

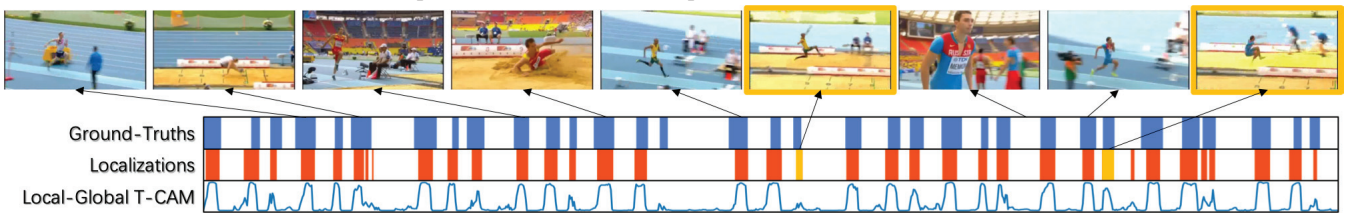
Action Recognition. We compare the action recognition performance of PreTrimNet with the state-of-the-art methods on THUMOS14 (Table 1) and ActivityNet1.3 (Table 2). As shown in the Tables, PreTrimNet remarkably outperforms its competitors in term of classification accuracy. Specifically, PreTrimNet with each single stream receives satisfactory results. When the three streams are fused, the accuracy can be further improved by integrating different aspects of the video contents. Considering that (to the best of our knowledge) we are the first to extract pose features for untrimmed video analysis, we report our results based on pose stream as a baseline for future reference.

Action Localization. We evaluate PreTrimNet for temporal action localization on THUMOS14 (Table 3) and ActivityNet1.3 (Table 4), in comparison with both fully and weakly supervised methods. It is observed that PreTrimNet significantly surpasses its weakly supervised counterparts. It is especially encouraging to see that PreTrimNet even achieves comparable results with some fully supervised methods.

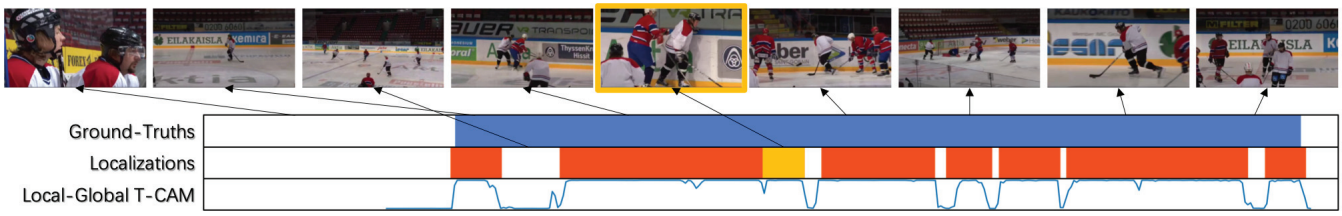
To validate the effectiveness of key components in PreTrimNet, we perform a set of ablation studies, comparing the full implementation of PreTrimNet with its abridged versions without one or both of pre-trimming and MIML clas-



(a) An example of $CricketBowling$ (top) and $CricketShot$ (down) actions.



(b) An example of $LongJump$ action.



(c) An example of $Playing\ ice\ hockey$ action.

Figure 3: Qualitative results on THUMOS14, numbered (a) and (b), and ActivityNet1.3, numbered (c). The horizontal axis denotes the timestamps. The red and yellow bars denote the time intervals detected with local and global T-CAM, respectively.

sification module. Table 5 summarize the action localization results on THUMOS14. We observe that each component is indispensable and makes its contribution to the localization performance.

We further illustrate examples of temporal localization on THUMOS14 and ActivityNet1.3, as shown in Figure 3, including (a) a video containing two action classes, (b) short-lasting action, and (c) long-lasting action. As we can see, the proposed local-global T-CAM is an effective indicator that is capable of locating actions of interest in untrimmed videos under different circumstances. It is noted that pose is a strong clue for human actions. The clips segmented based on pose successfully detect most frame-level actions with local T-CAM. Using global T-CAM as refiner, the localization performance can be further improved.

Conclusion

In this paper, we have proposed a novel framework, i.e. PreTrimNet, for effective action recognition and localization in untrimmed videos with video-level weak supervision. Using pose as an instructive guidance, person-centric clips that potentially contain actions of interest are extracted from untrimmed videos via spatio-temporal pre-trimming. The normalized pose sequences also serve as an additional stream to develop augmented feature representations. Under the multi-instance multi-label learning mechanism, the proposed method is capable of revealing the latent instance-label relation, which facilitates accurate identification of actions and the corresponding time intervals in untrimmed videos. As demonstrated on two challenging untrimmed video datasets, PreTrimNet achieves superior performance over the state-of-the-art weakly supervised methods, and is even comparable to some fully-supervised methods that leverage temporal annotations during training. In the future, inspired by the most recent advances in action recognition and localization, we will consider improving our framework by explicitly modeling backgrounds (Liu, Jiang, and Wang 2019), and further leveraging external sources (Nguyen, Ramanan, and Fowlkes 2019).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant 61871378).

References

Belagiannis, V., and Zisserman, A. 2017. Recurrent human pose estimation. In *FG*, 468–475.

Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 4724–4733.

Chao, Y.; Vijayanarasimhan, S.; Seybold, B.; Ross, D. A.; Deng, J.; and Sukthankar, R. 2018. Rethinking the faster R-CNN architecture for temporal action localization. In *CVPR*, 1130–1139.

Chen, X., and Yuille, A. L. 2015. Parsing occluded people by flexible compositions. In *CVPR*, 3945–3954.

Fang, H.; Xie, S.; Tai, Y.; and Lu, C. 2017. RMPE: regional multi-person pose estimation. In *ICCV*, 2353–2362.

Feng, J., and Zhou, Z. 2017. Deep MIML network. In *AAAI*, 1884–1890.

Ge, W.; Yang, S.; and Yu, Y. 2018. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *CVPR*.

Heilbron, F. C.; Escorcia, V.; Ghanem, B.; and Niebles, J. C. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 961–970.

Heilbron, F. C.; Barrios, W.; Escorcia, V.; and Ghanem, B. 2017. SCC: semantic context cascade for efficient action detection. In *CVPR*, 3175–3184.

Huang, S.; Gao, W.; and Zhou, Z. 2014. Fast multi-instance multi-label learning. In *AAAI*, 1868–1874.

Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; and Schiele, B. 2016. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 34–50.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 448–456.

Jiang, Y.-G.; Liu, J.; Roshan Zamir, A.; Toderici, G.; Laptev, I.; Shah, M.; and Sukthankar, R. 2014. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/.

Kiros, R.; Zhu, Y.; Salakhutdinov, R.; Zemel, R. S.; Urtasun, R.; Torralba, A.; and Fidler, S. 2015. Skip-thought vectors. In *NIPS*.

Lin, T.; Zhao, X.; Su, H.; Wang, C.; and Yang, M. 2018. BSN: boundary sensitive network for temporal action proposal generation. In *ECCV*, 3–21.

Liu, D.; Jiang, T.; and Wang, Y. 2019. Completeness modeling and context separation for weakly supervised temporal action localization. In *CVPR*.

Nguyen, P.; Liu, T.; Prasad, G.; and Han, B. 2018. Weakly supervised action localization by sparse temporal pooling network. In *CVPR*, 6752–6761.

Nguyen, P. X.; Ramanan, D.; and Fowlkes, C. C. 2019. Weakly-supervised action localization with background modeling. In *ICCV*.

Paul, S.; Roy, S.; and Roy-Chowdhury, A. K. 2018. W-TALC: weakly-supervised temporal activity localization and classification. In *ECCV*, 588–607.

Pham, A. T.; Raich, R.; Fern, X. Z.; and Arriaga, J. P. 2015. Multi-instance multi-label learning in the presence of novel class instances. In *ICML*, 2427–2435.

Pishchulin, L.; Jain, A.; Andriluka, M.; Thormählen, T.; and Schiele, B. 2012. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, 3178–3185.

Pishchulin, L.; Insafutdinov, E.; Tang, S.; Andres, B.; Andriluka, M.; Gehler, P. V.; and Schiele, B. 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*.

Sapp, B.; Toshev, A.; and Taskar, B. 2010. Cascaded models for articulated pose estimation. In *ECCV*, 406–420.

Shou, Z.; Chan, J.; Zareian, A.; Miyazawa, K.; and Chang, S. 2017. CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *CVPR*, 1417–1426.

Shou, Z.; Gao, H.; Zhang, L.; Miyazawa, K.; and Chang, S. 2018. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *ECCV*, 162–179.

Shou, Z.; Wang, D.; and Chang, S. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, 1049–1058.

Simonyan, K., and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 568–576.

Simonyan, K., and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Singh, K. K., and Lee, Y. J. 2017. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 3544–3553.

Su, H.; Zhao, X.; and Lin, T. 2018. Cascaded pyramid mining network for weakly supervised temporal action localization. In *ACCV*, 558–574.

Sun, M.; Kohli, P.; and Shotton, J. 2012. Conditional regression forests for human pose estimation. In *CVPR*, 3394–3401.

Toshev, A., and Szegedy, C. 2014. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 1653–1660.

Tran, D.; Bourdev, L. D.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 4489–4497.

Wang, F., and Li, Y. 2013. Beyond physical connections: Tree models in human pose estimation. In *CVPR*, 596–603.

Wang, H., and Schmid, C. 2013. Action recognition with improved trajectories. In *ICCV*, 3551–3558.

Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; and Xu, W. 2016a. CNN-RNN: A unified framework for multi-label image classification. In *CVPR*, 2285–2294.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Gool, L. V. 2016b. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 20–36.

Wang, L.; Xiong, Y.; Lin, D.; and Gool, L. V. 2017. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 6402–6411.

Xu, H., and Das, A. 2017. R-C3D: region convolutional 3d network for temporal activity detection. In *ICCV*, 5794–5803.

Yeung, S.; Russakovsky, O.; Mori, G.; and Fei-Fei, L. 2016. End-to-end learning of action detection from frame glimpses in videos. In *CVPR*, 2678–2687.

Yuan, J.; Ni, B.; Yang, X.; and Kassim, A. A. 2016. Temporal action localization with pyramid of score distribution features. In *CVPR*, 3093–3102.

Zhang, X.; Shi, H.; Li, C.; Zheng, K.; Zhu, X.; and Duan, L. 2019. Learning transferable self-attentive representations for action recognition in untrimmed videos with weak supervision. In *AAAI*, 9227–9234.

Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; and Lin, D. 2017. Temporal action detection with structured segment networks. In *ICCV*, 2933–2942.