Human Centric Spatio-Temporal Action Localization

Jianwen Jiang[1], Yu Cao[2], Lin Song[3], Shiwei Zhang[4] Yunkai Li[5], Ziyao Xu[5], Qian Wu[6], Chuang Gan[1*], Chi Zhang[5*], Gang Yu[5*]
[1]Tsinghua University, jjw17@mails.tsinghua.edu.cn, ganchuang1990@gmail.com
[2]Beihang University, cqcy1208@buaa.edu.cn
[3]Xian Jiaotong University, stevengrove@xtu.xjtu.edu.cn
[4]Huazhong University of Science and Technology, swzhang@hust.edu.cn
[5]Megvii Inc. (Face++), {liyunkai, xuziyao, zhangchi, yugang}@megvii.com
[6]Zhejiang University, wq1601@zju.edu.cn

***Abstract*—This paper describes our solution for the spatio-temporal action localization of ActivityNet AVA challenge. Our system is consisted of three components: a human detector, an action classification module and an actor-target relation network. We first apply a region proposal network (RPN) to detect human in the videos, since AVA mainly contains human-centric action categories. Then we conduct the action classification by adopting the ROI pooling operation on the human regions. In order to capture the human-object relationships, we further design an actor-target relation network, which is achieved with a non-local operation between the ROI and its surrounding regions. We finally obtains 25.63% and 25.75% in terms of mean average precision (mAP) on the validation set of the two tracks, and 21.075% and 20.99% on the testing set.**

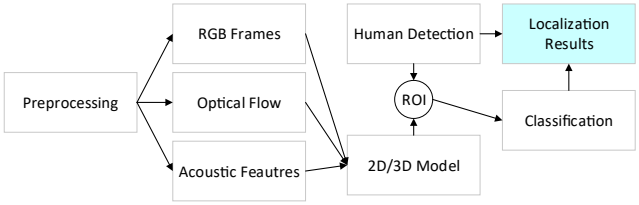

Fig. 1. The designed framework in our method. We split the spatio-temporal action localization into two subtasks, including human detection and action classification. Given the detections, we mainly focus on extracting multi vision cues, such as appearance information, motion information, and acoustic features. By applying ROI pooling, we can integrate the results from different models.

I. Introduction

Spatial-temporal action recognition and localization has received significant research attention in the computer vision communities [3], [28], [30], [31] due to its enormous applications such as public security, event recognition and video retrieval. There are some publicly available datasets such as UCF-Sports [16], J-HMDB [6] and UCF101 [21], [8], which have made great contribution to improve the performance for the task of action recognition and detection. Based on these benchmarks, there are a few promising deep model based methods, including TS (two streams) framework [18], C3D [22], TSN [25], p3d [14] and Artnet [23] for action recognition. These methods mainly try to extract different vision cues, such as short video clips [14], [22], [23], motion information [18] and long-range video clips [25]. Meanwhile, recent object detection methods, such as faster-RCNN [15], light head RCNN [9] and megdet [12], also make significant process for the general object detection.

Recently, some detection methods such as ACT [7], online method [20], multi-region faster-RCNN [13], achieve impressive results on the public datasets in the detection frameworks. In this challenge, the AVA dataset [4] is more challenging and we aim to apply different clues to extract video representation. In this report, we mainly adopt three modalities, including appearance, motion and audio information. Noting that, the audio feature is only applied in the full track. To conduct action detection, we design our method in the Faster-RCNN [15] framework. To better fit the framework on the action localization, we propose to apply a good pretrained human detectors as the RPN module, shown in Fig. I. Following the RPN, we train the action classification network in an end-to-end manner. Moreover, we design an actor-target relation (ATR) network to extract correlation between the actors and the corresponding targets. For this purpose, we conduct non-local operation between the ROI and its surrounding regions. The applied base models mainly focus on short- or long-term input clips, including i3d [1] with non-local module [26], C3D [22], and TSN [25].

For the RPN module, we apply FPN model [10] because of its high recall and precision. Given the proposal regions, we apply ROI pooling [15] to extract features and classify each proposals. After that, a posterior fusion strategy is used to give the final predictions of action categories of every corresponding target. Attributed to the structure of the designed model, we obtain about 10% gain than the baseline method [4]. We show the overview in the Fig. I.

The remaining sections are organized as follows. Section II presents the details of our method. In section IV, we also present some experimental results. Finally, this report concludes in section V.

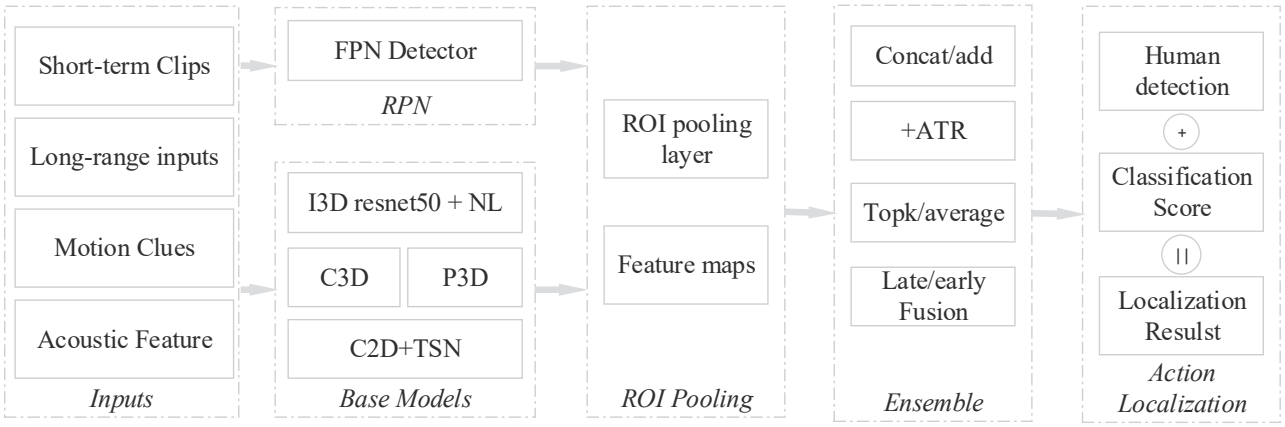

Fig. 2. The overview of our method. First, we explore different vision cues, which are respectively fed into RPN and feature extractors. Then we apply ROI pooling operation based on the proposal regions and the corresponding feature maps. After that, we explore different integration strategies on the applied models. Finally, we calculate the location results by considering the classification results and proposal regions.

II. The proposed Method

In this section, we first introduce the utilized multiple clues. Then we present the framework of action localization and classification in both tasks of AVA challenges.

III.

A. Multi clues

Action localization is a complex task and is very challenging. We explore several modalities for this task, including short-term clips, long-range temporal structure, motion information and acoustic features.

Short-term clips. Inspired by most 3D CNNs, such as C3D [22], P3D [14] and I3D [1], we apply several continuous clips as input to extract short-term video representation. As shown in the benchmark [4], long clip and large input size is helpful to improve the performance. Therefore, we explore the applied inputs with different length and size to better understand their influence to the final action classification results.

In the AVA dataset [4], the action instances are sparsely annotated per second. Therefore, we extract one clip in a short time interval to predict the target actions. In our method, we apply the I3D Resnet [26], and P3D [14] to conduct the video representation. All the models are pretrained in Kinetics [8] in advance.

Long-range sampling. TSN [25] is proved to be a powerful method of long-range temporal structure modeling. Similar to TSN [25], we apply uniform sampling strategy to sample n frames for the model to learn. In our method, we find it effective to apply the sampling in the traditional 2D CNN. Therefore, we just adopt 2D model in this framework, such as resnext [27], resnet [5] and artnet18 [24].

Before the ROI pooling module [15], we integrate the frame-level feature by using average pooling scheme along the time axis.

Motion clues. To better extract micro motion information between two consecutive frames, we calculate optical flow to be used as a modality of input for the deep models. We first compute the horizontal and vertical motion maps, which construct the two independent channels. For the third channel, we simply apply point-wise multiplication between these two maps.

In this paper, we extract optical flow by applying TV-L1 [29] method which is integrated in the Opencv tools. Moreover, we also explore different methods of optical flow, such as Farneback [2], to add variety to the modality.

Acoustic features. Acoustic information is also discriminative for some actions, such as "play musical instrument", "sing". Therefore, we try to extract acoustic feature to improve our video representation. Similar to CNN based audio classification task [17], we divide the videos into frames every 1s, following which Fourier transformation and histogram integration are adopted. Given the new frames, we apply a VGG16 [19] model to conduct action classification based on a pre-trained model on the Kinetics dataset.

B. Action Localization

In this section, we mainly introduce RPN module and classification module.

RPN module. The goal of these two tasks is to localize human centric spatio-temporal action, hence we hold the point that the RPN module should have good performance on human detection. In our method, we apply feature pyramid networks (FPN) [10] to reach this goal, for its better performance. The FPN detector is first pretrained on MSCOCO dataset [11] and then is fine tuned on the AVA dataset [4]. By this means, we can obtain 96.5% recall and 81.6% accuracy on the evaluation set on an Intersection over Union(IoU) threshold of 0.5.

Action classification. As aforementioned, we design our method in the faster RCNN [15] framework. We apply the ROI pooling [15] strategy based on the proposal regions and the corresponding base models. We locate the ROI pooling layer after the last feature maps, followed by a classification branch. For the classification network, sigmoid function is used as in [4]. Finally, the output of the classification branch

TABLE I
RESULTS ON VALIDATION SET.

Model	Input	Modality	Operation	mAP (%)
Faster-RCNN [4]	(3, 40(RGB)+40(Flow), 360, 400)	RGB + Flow	-	16.2
i3d resnet50 + NL	(3, 20, 224, 224)	RGB	-	19.33
	(3, 20, 224, 224)	RGB	ATR	20.01
	(3, 40, 224, 224)	RGB	40 clips	19.37
	(3, 20, 360, 400)	RGB	(360,400) size	19.86
	(3, 20(RGB)+20(Flow), 224, 224)	RGB + Flow	add	21.66
P3D199	(3, 20(RGB)+20(Flow), 224, 224)	RGB + Flow	-	17.87
resnet152	(3, 20, 224, 224)	RGB	TSN	14.68
artnet18	(3, 20, 224, 224)	RGB	-	16.67
Vgg16	-	Audio	-	6.5
Ensemble(Vison Only)				**25.63**
Ensemble (Full)				**25.75**

is used as the classification probability prediction results of the corresponding proposal boxes.

To further improve the performance, we also explore following several different strategies: *(i)* we concatenate or add the feature maps from different networks; *(ii)* we simply average the scores before or after sigmoid function; *(iii)* top-k fusion scheme are adopted for the ensemble process; *(iv)* we concatenate the features of RGB and Flow streams on the fully connected layer. *(v)* ROI align method is also explored.

In our method, we integrate all the model to calculate the results of our human detections. In the experiments, we find that apply $k = n//2$ (n is the number of the total applied models), and fusion before sigmoid function can lead to better results.

C. Extract Actor-Target Relationship

In the AVA dataset [4], we observed that the annotation boxes mainly contain the human but lose much attention on the targets, such as "grab (a person)" and "hug (a person)". We speculate the performance could be further improved by incorporating the the Actor-Target relationship (ATR).

Inspired by the successful application of the non-local [26] network on the action recognition, we adopt the non-local operation to extract ATR. Particularly, we conduct non-local operation between the ROI feature and features outside the bounding boxes. By this means, we can learn the discriminative relationship related to the actors. Experiments also show that this structure effectively captures the motion by bridging between people and the interactive objects through space and time domain.

D. Training

In this section, we present some details of our method during training stage. We train our network end-to-end with invariant 0.001 learning rate. For each model, we train about 5 epoches. We train our model on the 8 P40 GPUs for each experiments and the batch size is 16. When fusing different models, we freeze the base model before ROI pooling layer.

IV. EXPERIMENT RESULTS

In this section, we respectively report our performance on the validation and testing set in the Table I and Table II. In the Table I, we show the results with different 2D/3D models. All the 3D models are first pretrained on Kinetics [8], and all the 2D models are pretrained on the Imagenet. Extracting the ATR can obtain about 0.68%, which means it is indeed helpful to learning the relationship for action classification. Finally, our method obtain 25.63% and 25.75% in terms of mAP on the two tasks.

In the Table II, we finally get 21.075% and 20.99% mAP on the testing set. We find that there is a gap of about 5% between validation and testing set, we think the reason may be that there are different number of videos of the two sets. Noting that, the full track obtains inferior performance compared to computer vision (CV) only track on the test set, which it is opposite on the validation set. We think the reasons may include several following factors: (i) the acoustic features have little influence on the final results because of the just 6.5% mAP; (ii) when we submit our results to the service, we train our CV model on the training + validation set, while the acoustic model is just trained on the training set; (iii) the existing gap between the validation set and testing set weakens the contribution of our acoustic model.

TABLE II
RESULTS ON TESTING SET.

Tasks	mAP(%)
Computer Vision ONLY	21.075
Full	20.99

V. CONCLUSION

In the Activitynet-AVA Challenge 2018, we propose a new framework for the human centric spatio-temporal action localization. We design our method under the faster-RCNN framework, but propose to apply a good human detector as the RPN module. Meanwhile, we apply non-local operation between the proposal regions and their surrounding regions to extract actor-target relation (ATR). Moreover, we also explore different integration strategies to extract multi vision cues. By this means, we achieve significant improvement again the baseline method. In the future, we will explore the correlation between different actions and learn this correlation in the deep models.

REFERENCES

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

[2] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.

[3] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*, pages 2568–2577, 2015.

[4] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. *arXiv preprint arXiv:1705.08421*, 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199. IEEE, 2013.

[7] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid. Action tubelet detector for spatio-temporal action localization. In *ICCV*, 2017.

[8] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[9] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun. Light-head r-cnn: In defense of two-stage object detector. *arXiv preprint arXiv:1711.07264*, 2017.

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017.

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[12] C. Peng, T. Xiao, Z. Li, Y. Jiang, X. Zhang, K. Jia, G. Yu, and J. Sun. Megdet: A large mini-batch object detector. *arXiv preprint arXiv:1711.07240*, 2017.

[13] X. Peng and C. Schmid. Multi-region two-stream r-cnn for action detection. In *ECCV*, pages 744–759. Springer, 2016.

[14] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, 2017.

[15] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[16] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, pages 1–8. IEEE, 2008.

[17] e. S. Hershey, S. Chaudhuri. Cnn architectures for large-scale audio classificatio. *arXiv preprint arXiv:1609.09430*, 2017.

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

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR, abs/1409.1556*, 2014.

[20] G. Singh, S. Saha, M. Sapienza, P. Torr, and F. Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *ICCV*, pages 3637–3646, 2017.

[21] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

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

[23] L. Wang, W. Li, W. Li, and L. Van Gool. Appearance-and-relation networks for video classification. *arXiv preprint arXiv:1711.09125*, 2017.

[24] L. Wang, W. Li, W. Li, and L. Van Gool. Appearance-and-relation networks for video classification. *CVPR*, 2018.

[25] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool. Temporal segment networks: Towards good practices for deep action recognition. *ECCV*, 22(1):20–36, 2016.

[26] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. *arXiv preprint arXiv:1711.07971*, 2017.

[27] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 5987–5995. IEEE, 2017.

[28] G. Yu and J. Yuan. Fast action proposals for human action detection and search. In *CVPR*, pages 1302–1311, 2015.

[29] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223. Springer, 2007.

[30] S. Zhang, C. Gao, F. Chen, S. Luo, and N. Sang. Group sparse-based mid-level representation for action recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(4):660–672, 2017.

[31] S. Zhang, C. Gao, J. Zhang, F. Chen, and N. Sang. Discriminative part selection for human action recognition. *IEEE Transactions on Multimedia*, 20(4):769–780, 2018.