Action Recognition using Convolution Neural Network

Ming-Han Huang
Kuan-Chun Hong
cwhuangb1321@gmail.com
a0979478979@gmail.com
National Chiao Tung University
HsinChu, Taiwan

Figure 1: Recognition result of the scene in TV series: Hotel Del Luna

ABSTRACT

Localizing persons and recognizing their actions from videos has been a popular task in the field of computer vision and a challenging task in video understanding. The task can be divided into two subtasks: Object Detection and Action Recognition. Recent work shows that it can be achieved by recognize the actions in the ROIs. However, the relation between the actors and the background will affect the action results in the ROIs hardly. Therefore, we build a model that combines Yolov4 as the object detection model and I3D as the action recognition model and adopt Actor-Conditioned Attention Maps (ACAM) to take consider of the whole scene. The experiment results on AVA-Kinect dataset demonstrate the effectiveness of our work, results in 24.15 mAP.

CCS CONCEPTS

• **Computing methodologies** → **Object identification**.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

KEYWORDS

Spatio-Temporal Action Localization; Action Detection; Action Recognition; Relation; Attention

ACM Reference Format:
Ming-Han Huang and Kuan-Chun Hong. 2018. Action Recognition using Convolution Neural Network. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Spatio-temporal action localization, requiring localizing persons while recognizing their actions, is an important task that has drawn increasing attention in recent years [6, 8, 11, 27, 30]. It can make us more understand what is happening in the video. For example, it can be applied into the field of security. If the scenes from the monitors can be analyzing in the real-time, it can detect a crime scene immediately and call the police. Also, it can be applied in the field of elder care. A family member or the hospital can get the information of elder fall or emergency situation immediately and do the corresponding actions. Hence, we focus on the task of atomic action detection from videos in this work. We propose to model actor actions by using information from the surrounding context and evaluate our model on AVA-Kinect [11] dataset. We demonstrate the efficiency and transferability of our approach by implementing an action detection pipeline and qualitatively testing it on videos from various sources.

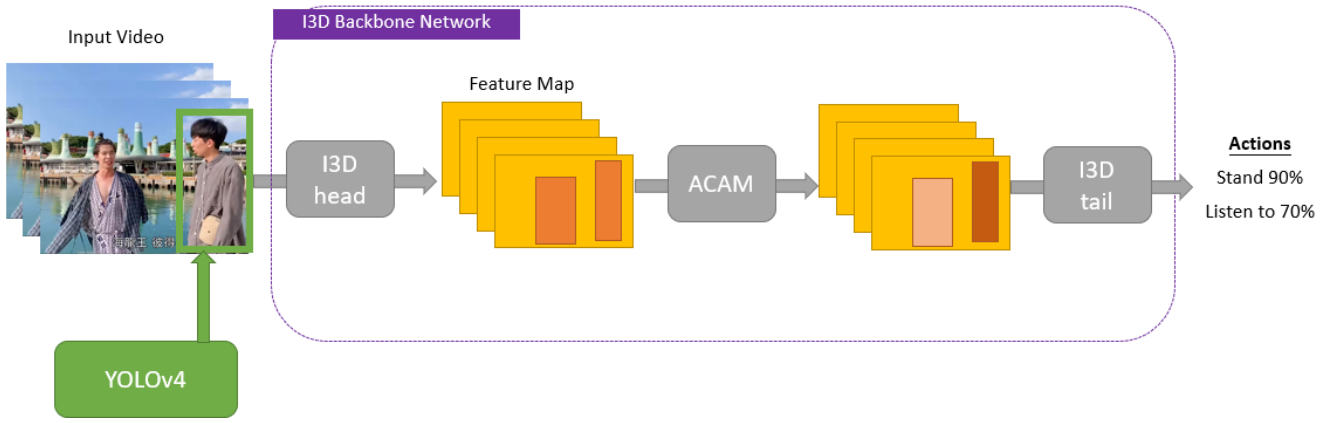

Figure 2: Proposed Model Solution

The task can be divided into two sub-tasks: Object Detection and Action Recognition. In the object detection part, we detect persons and objects from a video and return their bounding boxes. In this part, we do not know the action of each actor. The only information is the bounding boxes and the object's class. In the action recognition part, it will take in the whole video and tell us what is happening in the video. For example, we can feed a video clip of NBA plays and the model will tell us there is a basketball game in this video. However, this kind of model only tell us the overall situation of the video instead of the situation in the ROIs. Hence, we focus on integrating these two kinds of models into one model that can tell us what happened in the ROIs.

2 RELATED WORK

In the field of object detection, it is usually categorized into two kinds, i.e., one-stage object detector and two-stage object detector. The most representative two-stage object detector is the R-CNN[10] series, including fast R-CNN[9], faster R-CNN[24], R-FCN[4], and Libra R-CNN[20]. It is also possible to make a two stage object detector an anchor-free object detector, such as RepPoints[32]. As for one-stage object detector, the most representative models are YOLO[1, 21–23], SSD[18], and RetinaNet [17].

Early works mainly focus on classifying a short video clip into an action class. 3D-CNN[2, 28], two-stream network[26, 29] and 2D-CNN with RNN[5, 19] are the three dominant network architectures adopted for this task. While progresses are made for short trimmed video classification, the main research stream moves forward to understand long untrimmed videos, which requires not only to recognize the category of each action instance but also to locate its start and end times. A handful of works[25] consider this problem as a detection problem in 1D temporal dimension by extending from object detection frameworks. The release of the large-scale, high quality datasets like Sports 1M[13], Kinetics[14], allowed deeper 3D CNN models such as C3D[31], Inception 3D (I3D)[3] to be trained and achieve higher performance.

Recently, the problem of spatio-temporal action localization has drawn considerable attention of the research community, and datasets such as AVA, where atomic actions of all actors in the video are continuously annotated, are introduced. It brings the action detection problem into a finer level, since the action instance needs to be localized in both space and time. Typical approaches used by early works adopted R-CNN detectors for object detection on 3D-CNN features[11]. Several more recent works have exploited graph-structured networks to leverage contextual information [8, 27, 34]. In particular, some approaches utilize the self-attention mechanism to learn relationships among actors. Among them, Wu et al.[30] proposed to use long-term feature banks (LFB) to provide temporal supportive information up to 60s; ACRN[27] models relations between human actors and scene elements through a relation network. Our proposed model leverages this relation idea to generate attention maps.

3 PROPOSED SOLUTION

In this section, we will describe our proposed model for action recognition. The goal is to detect bounding boxes for each actor and classify their actions from each input video segment. Each actor can have multiple action labels, such as "sitting" and "talk to", simultaneously.

First, the Yolov4 processes the input video segments. The reason we choose Yolov4 as our object detection model is its performance in real-time detection. It will detect the objects in the input video segments and then return the boundary boxes and the classes of the objects. We will filter these boundary boxes and remain the boundary boxes that for persons. Next, the input video segments are processed by the I3D back-bone. Feature vectors for each detected actor are generated from their locations on the feature map. In ACAM, a set of weights is generated for every spatio-temporal region in the scene by combining the actor features and contextual features extracted from the entire scene. These weights are multiplied by the feature map and the result represents the actor conditioned features. For example, in Fig.2, two detected actors are represented by two vertical bars in feature map. One focused actor (boxed) is listening to a close-by actor. This action is captured by larger weights in the attention map shown as a darker vertical bar. In the end, the generated actor conditioned features is then

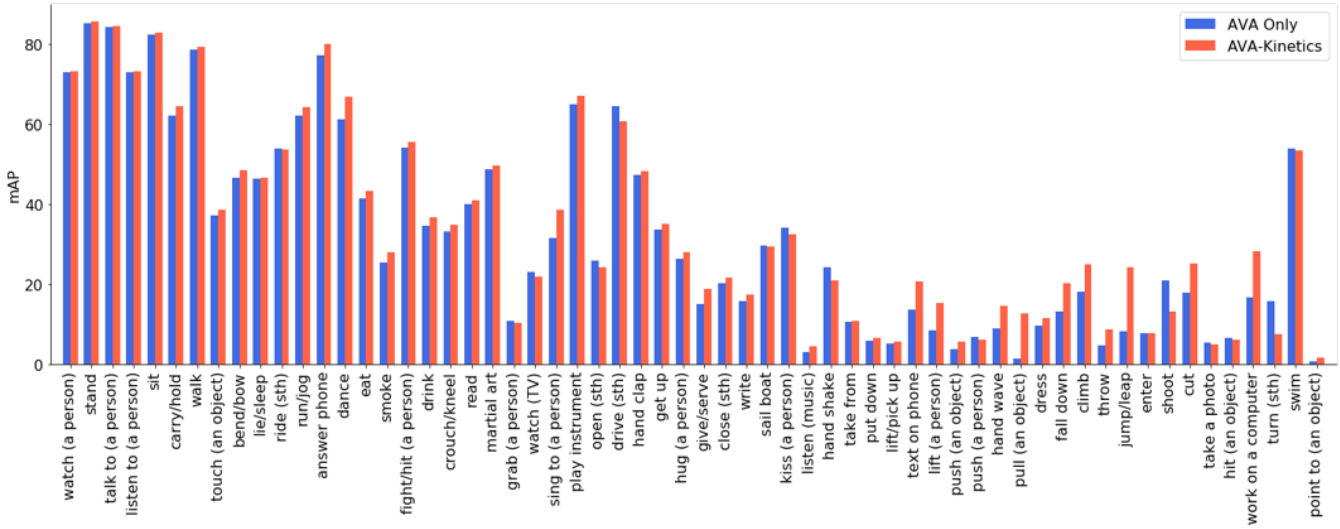

Figure 3: Impact of Kinect-700.

classified by remaining layers of the CNN back-bone. The proposed model architecture is shown in Fig.2.

4 EXPERIMENT

In this section, we evaluate our proposed model on the challenging AVA dataset[11]. We first introduce some implementation details.

4.1 Datasets & Implementation Details

Dataset. AVA[11] is a video dataset of spatio-temporally localized atomic visual actions. In addition to the current AVA dataset, Kinetics-700[2] videos with AVA style annotations are recently introduced. The new AVA-Kinetics dataset[15] of spatio-temporally localized atomic visual actions contains over 238k unique videos and more than 624k annotated frames. For AVA, box annotations and their corresponding action labels are provided on key frames of 430 15-minute movie clips with a temporal stride of 1 second, while for Kinetics only a single frame is annotated for each video. The AVA dataset is challenging since movie scenes are often highly complex and contain multiple actors, each of which may perform several atomic actions simultaneously. Following the guidelines of the AVA benchmark, we evaluate on 60 action classes, and the performance metric is mean Average Precision (mAP) using a frame-level IoU threshold of 0.5.

Person Detector. As for person detection on key frames, we use pre-computed human bounding box proposals from Yolov4[1].

3D CNNs We use I3D[3] as the 3D CNN back-bone for all of our model candidates. The input video segment of RGB frames is processed by the initial I3D layers until the layer to obtain the feature tensor. The actor conditioned features are computed using ACAM. The remaining I3D layers are used and initialized with pre-trained weights. We use the remaining layers up to final for classification on the actor conditioned features and call this operation "I3D Tail". A global average pooling across spatio-temporal dimensions is applied to the final feature map to compute class probabilities.

4.2 Experiment Results

We compare our results with state-of-the-art methods on the AVA validation set in Table 1. With more advanced I3D backbone, our model reaches 22.29 mAP on AVA v2.1, surpassing all prior results. On the other hand, with AVA v2.2 and finer pre-training with Kinetics-700, our model achieves 24.15 mAP with only single-scale testing, establishing a new state-of-the-art on AVA.

Table 1: Validation mAP results compared to published state of the art results.

Model	mAP
I3D[11]	15.6
ACRN[27]	17.4
YH Technologies[33]	19.4
Megvii/Tsinghua[12]	20.01
Deep Mind[7]	21.9
Ours(AVA only)	22.29
RTPR[16]	22.3
Ours(AVA-Kinetics)	24.15

4.3 Discussions

From Fig 3 we can see the mAP of each categories. The mAP of the regular actions such as "watch (a person)", "stand" or "talk to" have very high mAP comparing to the more complex actions such as "pull (an object)" or "listen (music)". The reason might be the data imbalance. Since the actions like "stand" or "sit" will definitely appear in every action, they have more training data and it's like a binary classification for the model: a person must be now standing or sitting. This is slightly improved after we add the data from Kinect-700, give us more data of complex actions.

Figure 4: Sample results. Left person: walk, carry/hold; Right person: walk, carry/hold.

Figure 5: Sample results. Left person: sit, carry/hold, watch, talk to, listen to; Right person: sit, drive, listen to.

5 CONCLUSION

Given the high complexity of realistic scenes encountered in the spatio-temporal aciton localization task which involve multiple actors and a large variety of contextual objects, we observe the demand for a more sophisticated form of relation reasoning than current ones which often miss important hints for recognizing actions. Therefore, we introduce the concept of integrating three state-of-the-art models: Yolov4, I3D and ACAM, which uses attention maps as a set of weights to highlight the spatio-temporal regions that are relevant to the actor, while damping irrelevant ones. Extensive experiments on the action localization task show our model leads to a significant performance gain and achieves state-of-the-art results on the challenging AVA dataset.

REFERENCES

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934 [cs.CV]

[2] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A Short Note on the Kinetics-700 Human Action Dataset. arXiv:1907.06987 [cs.CV]

[3] Joao Carreira and Andrew Zisserman. 2018. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. arXiv:1705.07750 [cs.CV]

[4] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. 2016. R-FCN: Object Detection via Region-based Fully Convolutional Networks. arXiv:1605.06409 [cs.CV]

[5] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. 2016. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. arXiv:1411.4389 [cs.CV]

[6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. arXiv:1812.03982 [cs.CV]

[7] Rohit Girdhar, João Carreira, Carl Doersch, and Andrew Zisserman. 2018. A Better Baseline for AVA. arXiv:1807.10066 [cs.CV]

[8] Rohit Girdhar, João Carreira, Carl Doersch, and Andrew Zisserman. 2019. Video Action Transformer Network. arXiv:1812.02707 [cs.CV]

[9] Ross Girshick. 2015. Fast R-CNN. arXiv:1504.08083 [cs.CV]

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv:1311.2524 [cs.CV]

[11] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. 2018. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. arXiv:1705.08421 [cs.CV]

[12] Jianwen Jiang, Yu Cao, Lin Song, Shiwei Zhang, Yunkai Li, Ziyao Xu, Qian Wu, Chuang Gan, Chi Zhang, and Gang Yu. 2018. Human Centric Spatio-Temporal Action Localization.

[13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 1725–1732. https://doi.org/10.1109/CVPR.2014.223

[14] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950 [cs.CV]

[15] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. 2020. The AVA-Kinetics Localized Human Actions Video Dataset. arXiv:2005.00214 [cs.CV]

[16] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. 2018. *Recurrent Tubelet Proposal and Recognition Networks for Action Detection: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VI.* 306–322. https://doi.org/10.1007/978-3-030-01231-1_19

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal Loss for Dense Object Detection. arXiv:1708.02002 [cs.CV]

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. *Lecture Notes in Computer Science* (2016), 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

[19] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond Short Snippets: Deep Networks for Video Classification. arXiv:1503.08909 [cs.CV]

[20] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. 2019. Libra R-CNN: Towards Balanced Learning for Object Detection. arXiv:1904.02701 [cs.CV]

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640 [cs.CV]

[22] Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Better, Faster, Stronger. arXiv:1612.08242 [cs.CV]

[23] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. arXiv:1804.02767 [cs.CV]

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497 [cs.CV]

[25] Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. arXiv:1601.02129 [cs.CV]

[26] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. arXiv:1406.2199 [cs.CV]

[27] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. 2018. Actor-Centric Relation Network. arXiv:1807.10982 [cs.CV]

[28] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. arXiv:1412.0767 [cs.CV]

[29] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. arXiv:1608.00859 [cs.CV]

[30] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross Girshick. 2019. Long-Term Feature Banks for Detailed Video Understanding. arXiv:1812.05038 [cs.CV]

[31] Huijuan Xu, Abir Das, and Kate Saenko. 2017. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. arXiv:1703.07814 [cs.CV]

[32] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. 2019. RepPoints: Point Set Representation for Object Detection. arXiv:1904.11490 [cs.CV]

[33] Ting Yao and Xue Li. 2018. YH Technologies at ActivityNet Challenge 2018. arXiv:1807.00686 [cs.CV]
[34] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. 2019. A Structured Model For Action Detection. arXiv:1812.03544 [cs.CV]

A GITHUB LINK

https://github.com/hannnnk1231/109-1-Data-Science-Project