This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

An Effective Temporal Localization Method with Multi-View 3D Action Recognition for Untrimmed Naturalistic Driving Videos

Manh Tung Tran[§], Minh Quan Vu[§], Ngoc Duong Hoang, Khac-Hoai Nam Bui[*]
Viettel Cyperspace Center, Viettel Group
Hanoi, Vietnam

Abstract

Naturalistic driving studies with computer vision techniques have become an emergent research issue. The objective is to classify the distracted behavior actions by drivers. Specifically, this issue is regarded as temporal action localization (TAL) of untrimmed videos, which is a challenging task in the research field of video analysis. Particularly, TAL remains as one of the most challenging unsolved problems in computer vision that requires not only the recognition of action but the localization of the start and end times of each action. Most state-of-the-art approaches adopt complex architectures, which are expensive training and inefficient inference time. In this study, we propose a new framework for untrimmed naturalistic driving videos by utilizing the results from 3D action recognition with video clip classification for short temporal and spatial correlation. Then, simple post-processing based on data-driven is presented for long temporal correlation in untrimmed videos. The proposed method is evaluated on the AI City Challenge 2022 dataset for Naturalistic Driving Action Recognition. Accordingly, our method achieves the top 1 on the public leaderboard of the challenge.

1. Introduction

Video analysis is an important process for developing various applications such as robotics, human-computer interaction, and intelligent surveillance. Recently, the research on applying video analysis for intelligent transportation systems (ITS) has been paid more attention due to the rapid development of deep learning (DL) models for detection and recognition [1, 10]. In the domain of ITS, video driver behavior analysis is becoming one of the most important tasks for intelligent vehicles [11]. Specifically, naturalistic driving studies serve as an essential tool in studying driver behavior in real-time, which capture the action of the driver in traffic environments. However, the lack of labels and poor quality of data make it difficult to apply this study in practical. In this regard, AI City Challenge has recently published a new dataset and organized a competition of naturalistic driving action recognition [1]. Accordingly, the synthetic naturalistic data has been collected from multiple cameras inside the vehicle and the objective is to classify the distracted behavior activities by the driver in a given time frame. Technically, this task includes two main technical challenges as follows:

- The video classification system should be able to recognize activities in untrimmed videos, which include multiple actions of drivers. To the best of our knowledge, the provided dataset in this task obtains the most number of actions (labels) in terms of naturalistic driving studies with high appearance similarities among driver's actions.

- The final output should include temporal segments in which the actions appear in the video. Specifically, most recent state-of-the-art models rely on complex end-to-end architectures with large-scale dataset to train temporal localization models, which make those systems are difficult to be applied in practical.

In order to deal with both aforementioned challenges, in this study, we present an effective framework for naturalistic driving studies. Specifically, a 3D action recognition based on X3D (Expanding 3D) architecture with multi-view processing is employed for spatial and short temporal correlations. Then, a simple post-processing is presented to localize the long temporal correlations of untrimmed videos. The general pipeline of the proposed method is illustrated in Fig. 1.

[*]Corresponding email: hoainam.bk2012@gmail.com
[§]Equal contribution

[1]https://www.aicitychallenge.org/2022-challenge-tracks/. Accessed by April, 7th 2022

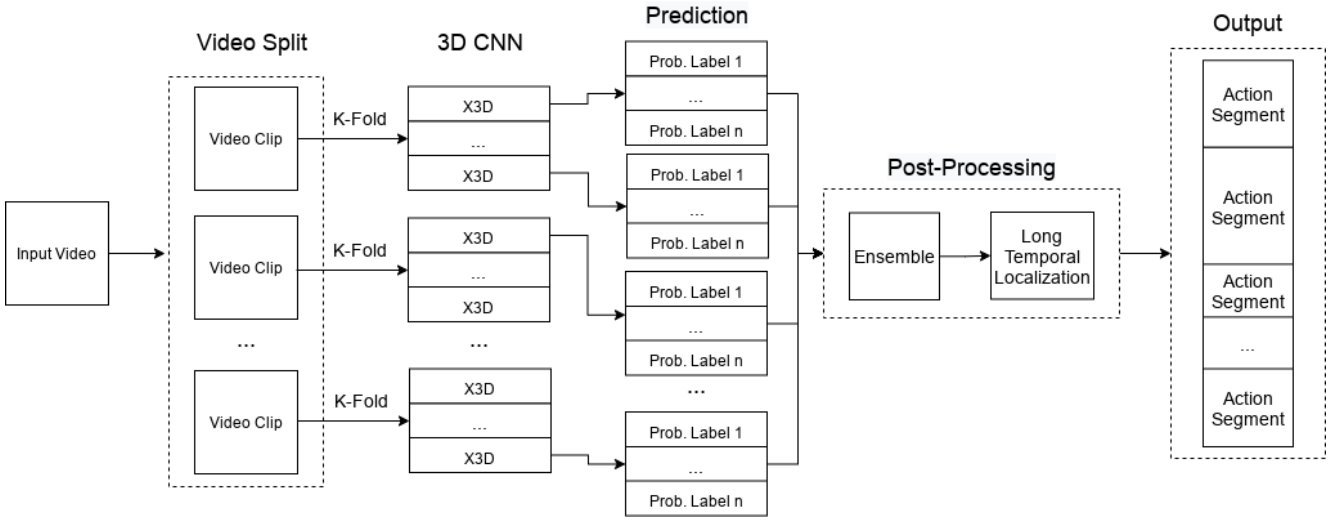

Figure 1. The general pipeline of proposed framework. Each input video is split into multiple video clips (the number of video clips depends on the length of the input video). Each video clip is put into X3D model for extracting short temporal and spatial correlation. Furthermore, ensemble technique for K models is adopted for improving the performance. Sequentially, the output is put into post-processing with simple methods for extracting long temporal correlation.

2. Background

2.1. Temporal Action Localization

Current methods for TAL are technically categorized into different approaches, which depends on the pipelines such as: i) *Multi-stage* methods perform frame (or segment) level classification with post-processing for obtaining temporal boundaries of actions [4, 14]; ii) *Two-stage* methods, similar with two-stage object detection in images, are technically a special type of multi-stage with one-stage proposal generator by directly predicting the scores and boundaries with each temporal location [12,16]; iii) *One-stage* methods has recently been proposed by integrating proposal generation and classification into end-to-end architectures [5, 15].

Specifically, one-stage methods with end-to-end manner have become an emerging approach in this research field. However, this approach has to face with the computational cost in terms of both complexity architectures and large training datasets [13]. Therefore, our solution utilizes the concept of both two-stage and one-stage paradigms by dividing the input video into multiple clips and process them separately using 3D action recognition models. In particular, this method is able to provide highly reliable video clip classifier.

2.2. X3D for Video Action Recognition

In this study, we adopt X3D [2], a state-of-the-art spatial-temporal (3D) network for detecting the action in each video clip. Technically, comparing with 2D CNN, 3D CNN contains more parameters, which lead the problem of computational heavy. X3D network has been introduced for reducing the number of parameters by expending an axis from a tiny spatial network (*e.g., space, time, width, and depth*). Accordingly, there are total six variant models of X3D, which range from extra small (XS) to extra extra large (XXL) (*i.e., X3D-XS, X3D-S, X3D-M, X3D-L, X3D-XL, and X3D-XXL*) based on the complexity regimes by FLOPs [9].

3. Proposed Framework

3.1. 2022 AI City Challenge Dataset

Regarding the Track 3 of Naturalistic Driving Action Recognition, the 2022 AI City Challenge provides a synthetic distracted driving (SynDD1) dataset, which have collected from a stationary vehicle using three in-vehicle cameras positioned at locations such as on the Dashboard, near the Rearview mirror, and on the top Right-side window corner [8]. Specifically, the dataset contains 90 video clips (about 14 hours in total) capturing 15 drivers with total 18 actions as shown in Tab. 1. The 14 hours of videos in this track are split into three datasets including A1 for training, A2 and B for testing (5 drivers for each dataset). The training dataset obtains the ground truth labels of start time, end time and types of actions. Accordingly, due to the complex similarities among actions and overlapping views of input videos, our proposed framework adopts ensemble technique with multi-view processing to improve the performance. More details of this process are described in the following section.

ID	Description	ID	Description
0	Normal Driving	9	Adjust control panel
1	Drinking	10	Pick up from floor (Dri.)
2	Phone Call(right)	11	Pick up from floor (Pax)
3	Phone Call(left)	12	Talk to Pax (right)
4	Eating	13	Talk to Pax (backseat)
5	Text (right)	14	Yawning
6	Text (left)	15	Hand on head
7	Hair / makeup	16	Singing
8	Reaching behind	17	Shaking or Dancing

Table 1. 18 distracted behavior actions of SynDD1. Label 0 is not considered for the evaluation.

3.2. Ensemble Model with Multi-View Processing

We use X3D Large (X3D-L) as the action recognition network with 5 Fold cross validation (K = 5). Since the dataset provides three views for driver's actions (*i.e., Dashboard, Rear View, and Right Side*), we modified the general pipeline by adding the ensemble technique with multi-view processing as shown in Fig. 2. Accordingly, the final prediction extracted from 15 single models by using average scores is generally formulated as follows:

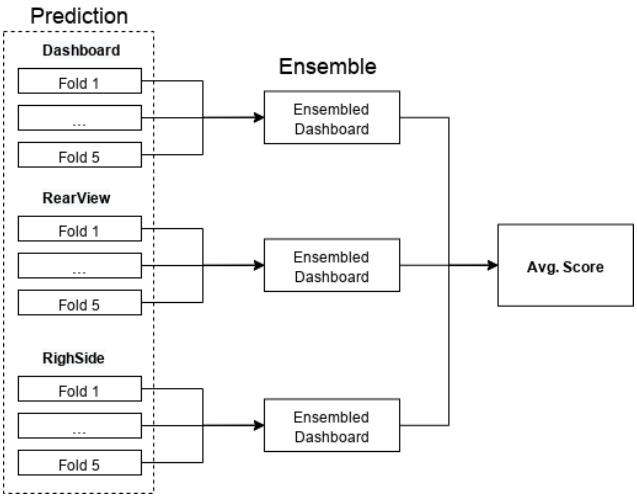

Figure 2. Ensemble with multi-view processing for the prediction.

$$Z = mean(E_1 + E_2 + ... + E_v) \quad (1)$$

where v denotes the number of camera views and $S_i = mean(Score_{K-Fold})$ represents the ensemble scores of camera views i.

3.3. Post-Processing

Given an input video, the output predictions are probability scores of all actions in each video clip. The outputs are then post-processed for predicting the action label and temporal localization of the predicted action. Technically, we consider the class with maximum probability score as the predicted class. Normally, Non-maximum suppression (NMS) algorithm [7] is widely used for the post-processing of TAL problem, which is able to remove redundant proposal and achieve higher recall with fewer proposals. However, based on our observation, this well-known post-processed algorithm might not be suitable with the AI City Challenge dataset because the ground-truth labels of actions are not overlapped in each video clip. Therefore, in this study, we adopt two other post-processing methods for this challenge.

The first method follows the work in [6]. Specifically, the smoothing filter process is adopted to smooth the values using mean filter, which is formulated as follows:

$$\hat{P}_l(x) = \frac{1}{2w} \sum_{l-w}^{l+w} p_l(x) \quad (2)$$

where w is the window size. p(x) and l(x) denote the sequence of probability scores and its length, respectively. Sequentially, the action is predicted based on the new probabilities scores for each video clips. Then, the predicted action of each video clip is labelled with previously predicted label if the probability score overs a threshold value.

The second method is our simple custom post-processing method based on the characteristic of AI City Challenge Dataset. Specifically, we first filter the low-score predicted label into none label. The threshold value of this step (λ_1) is determined by average value of probability scores. The long temporal correlation of an action (action segment) is then processed by merging video clips with same labels. In this regards, two clips with same class are merged if their temporal correlation is smaller than a threshold value (λ_2, by second). The last step is to remove the action segments of which the total time is smaller than a constant value (λ_3, by second). Notably, λ_2 and λ_3 are hyperprameters, which are tuned during the inference. Fig. 3 illustrates the main steps of our custom post-processing method.

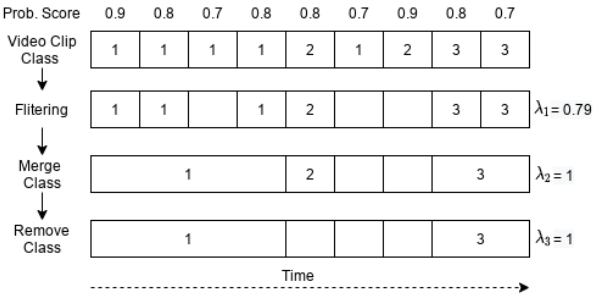

Figure 3. An output example using the custom post-processing for TAL of untrimmed videos.

3.4. Inference Complexity

The complexity is linear to the number of fold (K) and camera views (v). Specifically, the computational complexity of the proposed framework is formulated as follows:

$$O_{ensemble} = K * v * O_{X3D-L} \tag{3}$$

where O_{X3D-L} denotes the computational complexity of a single X3D-L model.

4. Experiment

4.1. Experiment Setting

Model configuration: We employ X3D-L model, in which the operations of width and depth are 2.0, and 5.0, respectively. The model is developed based on PySlowFast [2]. The pretrained model on Kinetics dataset, is loaded from PySlowFast library.

Training configuration: The training dataset is processed following the format of Kinetics dataset, in which the training video is divided into multiple segments [3]. The model is fine-tuned on AI City Challenge dataset with Adam optimizer for 18 epochs. Learning rate is initialized as 1e-6 and decreased by cosine schedule with the learning rate 5e-4. The scale Jittering during training is randomly selected from 512 to 640, and training crop size is 448. For the temporal domain, different with the original model, instead of one sample, two random sample/clips from the input segment are used for each training epoch. The number of frames and sampling rate are set to 8 and 4, respectively. Furthermore, we set all video clips without any label to label 0 (normal forward driving). All experiments are executed with batch size 8 on a single Nvidia A100 GPU.

Inference: During the inference, the value of k for smoothing is set to 3. The threshold values of λ_2 and λ_3 are set to 16 and 4 seconds, corresponding to number of 16 and 4 video clips, respectively.

4.2. Metrics

The evaluation uses F1 score as the metrics, which can be calculated as follows:

$$\begin{aligned} F1\ Score &= 2 \times \frac{precision \times recall}{precision + recall} \\ &= \frac{TP}{TP + 1/2(FP + FN)} \end{aligned} \tag{4}$$

where TP (True-Positive) represents objects correctly identified within the region of interest. FP (False-Positive) denotes the identified objects that are not TP identification. FN (False-Negative) identification is a ground-truth object that are not correctly identified. The final results of F1 Score is calculated by the evaluation portal of 2022 AI City Challenge [3].

4.3. Results Analysis

K-Fold: 5-Fold is applied for cross validation corresponding to 5 drivers of training dataset in which the data of 04 drivers are used for training and other driver data for validation. Tab. 2 show the results of training process for each fold with three views, respectively. Accordingly, results are significantly different when we change the validation set of user (driver) data, which lead to the hypothesis that the model might not be stable if we specify a permanent user for the validation. Furthermore, the results also change following the camera views with different user id. For instance, the user id 49381 is able to achieve the high accuracy in Right Side of camera view but low performance in the Dashboard view, opposite to user id 3513. Therefore, ensemble technique for all aforementioned models is needed to improve the performance of the challenge.

Cam. Views	Fold	Val user id	Epoch	Accuracy
Dashboard	1	24026	16	77.78
	2	24491	13	62.86
	3	35133	05	70.59
	4	38058	10	44.12
	5	49381	10	58.33
Rear View	1	24026	13	69.44
	2	24491	13	60.00
	3	35133	13	64.71
	4	38058	15	52.94
	5	49381	08	61.11
Right Side	1	24026	16	80.56
	2	24491	10	54.29
	3	35133	09	38.24
	4	38058	08	47.06
	5	49381	06	69.44

Table 2. Results of training process of K-Fold cross-validation

Post-Processing: The results of two post-processing methods presented in this study are shown in Tab. 3. The experiment executes with two camera views (*i.e., Dashbroad and Right Side*). The number frames and sampling rate are set to 15 and 4, respectively. In general, the second method is slightly better than the first method. However, the precision of the first method is better, which leads to a hypothesis that the performance of post-processing can be improved by combining both methods. We leave this issue as the future work of this study. For the final ranking, which is tested on the A2 dataset, we use the second method in the

[2] https://github.com/facebookresearch/SlowFast

[3] https://eval.aicitychallenge.org/aicity2022/

Method	F1 Score	Precision	Recall
First Method	0.2379	0.2803	0.2067
Second Method	0.2567	0.2756	0.2402

Table 3. Comparison results between two post-processing methods

post-processing phase to extract the long temporal action correlation.

4.4. Final Ranking

Tab. 4 shows the top teams from the public leader board of the challenge. Our proposed framework achieved at the

Rank	Team ID	Score
1	**72 (Our)**	**0.3492**
2	43	0.3295
3	97	0.3248
4	15	0.3154
5	78	0.2921
6	16	0.2905
7	106	0.2902
8	124	0.2849
9	54	0.2710
10	95	0.2706

Table 4. Top 10 Leaderboard of AIC Challenge 2022 Track 3 Naturalistic Driving Action Recognition

first place with 0,3492 of F1 Score. More detailed results in terms of precision and recall scores are shown in Tab. 5

F1 Score	Precision	Recall
0.3492	0.4044	0.3073

Table 5. Detailed Results of the proposed Framework

5. Conclusion

This study presents a solution for AI City Challenge 2022 in Track 3, which focuses on temporal action localization for Naturalistic Driving videos. Specifically, the proposed framework includes two phases: i) The first phase adopts 3D action recognition by using X3D Large model for extracting short temporal and spatial correlation; ii) The second phase presents a post-processing method for localizing long temporal correlation. The experiment on A2 dataset with around 10 input videos shows that the proposed framework achieves at the first place of the challenge, in which the F1 Score, Precision, and Recall are 0,3492, 0.3073, and 0.4044, respectively.

References

[1] Khac-Hoai Nam Bui, Hongsuk Yi, and Jiho Cho. A vehicle counts by class framework using distinguished regions tracking at multiple intersections. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 2466–2474. Computer Vision Foundation / IEEE, 2020. 1

[2] Christoph Feichtenhofer. X3D: expanding architectures for efficient video recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 200–210. Computer Vision Foundation / IEEE, 2020. 2

[3] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 4

[4] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: boundary-matching network for temporal action proposal generation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3888–3897. IEEE, 2019. 2

[5] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *CoRR*, abs/2106.10271, 2021. 2

[6] Alberto Montes, Amaia Salvador, and Xavier Giró-i-Nieto. Temporal activity detection in untrimmed videos with recurrent neural networks. *CoRR*, abs/1608.08128, 2016. 3

[7] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR 2006), 20-24 August 2006, Hong Kong, China*, pages 850–855. IEEE Computer Society, 2006. 3

[8] Mohammed Shaiqur Rahman, Archana Venkatachalapathy, Anuj Sharma, Jiyang Wang, Senem Velipasalar Gursoy, David Anastasiu, and Shuo Wang. Synthetic distracted driving (syndd1) dataset for analyzing distracted behaviors and various gaze zones of a driver. *CoRR*, 2022. 2

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2

[10] Shaohua Wan, Xiaolong Xu, Tian Wang, and Zonghua Gu. An intelligent video analysis method for abnormal event detection in intelligent transportation systems. *IEEE Trans. Intell. Transp. Syst.*, 22(7):4487–4495, 2021. 1

[11] Yang Xing, Chen Lv, Huaji Wang, Dongpu Cao, Efstathios Velenis, and Fei-Yue Wang. Driver activity recognition for intelligent vehicles: A deep learning approach. *IEEE Trans. Veh. Technol.*, 68(6):5379–5390, 2019. 1

[12] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: region convolutional 3d network for temporal activity detection. In *IEEE International Conference on Computer Vision, ICCV*

2017, Venice, Italy, October 22-29, 2017, pages 5794–5803. IEEE Computer Society, 2017. 2

[13] Mengmeng Xu, Juan-Manuel Pérez-Rúa, Victor Escorcia, Brais Martínez, Xiatian Zhu, Li Zhang, Bernard Ghanem, and Tao Xiang. Boundary-sensitive pre-training for temporal localization in videos. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 7200–7210. IEEE, 2021. 2

[14] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali K. Thabet, and Bernard Ghanem. G-TAD: sub-graph localization for temporal action detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10153–10162. Computer Vision Foundation / IEEE, 2020. 2

[15] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. *CoRR*, abs/2202.07925, 2022. 2

[16] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. *Int. J. Comput. Vis.*, 128(1):74–95, 2020. 2