Multi-View Action Recognition for Distracted Driver Behavior Localization

Yuehuan Xu[1]*, Shuai Jiang[1]*, Zhe Cui[1,2]†, Fei Su[1,2]
[1]Beijing University of Posts and Telecommunications
[2]Beijing Key Laboratory of Network System and Network Culture, China
{xuyuehuan,js,cuizhe,sufei}@bupt.edu.cn

Abstract

The detection and recognition of distracted driving behaviors has emerged as a new vision task with the rapid development of computer vision, which is considered as a challenging temporal action localization (TAL) problem in computer vision. The primary goal of temporal localization is to determine the start and end time of actions in untrimmed videos. Currently, most state-of-the-art temporal localization methods adopt complex architectures, which are cumbersome and time-consuming. In this paper, we propose a robust and efficient two-stage framework for distracted behavior classification-localization based on the sliding window approach, which is suitable for untrimmed naturalistic driving videos. To address the issues of high similarity among different behaviors and interference from background classes, we propose a multi-view fusion and adaptive thresholding algorithm, which effectively reduces missing detections. To address the problem of fuzzy behavior boundary localization, we design a post-processing procedure that achieves fine localization from coarse localization through post connection and candidate behavior merging criteria. In the AICITY2024 Task3 TestA, our method performs well, achieving Average Intersection over Union(AIOU) of 0.6080 and ranking eighth in AICITY2024 Task3. Our code will be released in the near future.

1. Introduction

Distracted driving is highly dangerous for human life. At present, naturalistic driving studies and computer vision techniques provide the much needed solution to identify and eliminate distracting driving behavior on the road [25]. Its key technologies mainly involved action recognition and temporal action localization methods in computer vision. Temporal action localization is aimed at identifying the start and end time of actions in untrimmed videos.

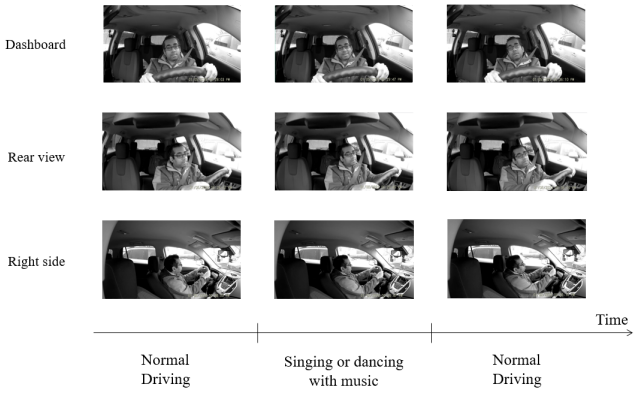

Figure 1. An challengeing example of the synthetic naturalistic data of the AI City Challenge 2024.

However, lack of labels, poor data quality and low resolution have created obstacles for deriving insights from data pertaining to the driver in the real world. Fortunately, 2024 AI City challenge Track 3, as shown in Fig. 1, provides high-quality datasets collected from multiple cameras at different positions inside the vehicle, which has facilitated further research in natural driving behavior. Technically, this task can be categorized as a temporal localization problem. The main challenge is to require the system to accurately detect and identify behaviors in untrimmed videos, which may contain multiple segments of behaviors, and output behavior labels, start and end times in untrimmed videos. While the similarity of different behaviors is high, which makes it difficult to distinguish; The boundaries of behavior are blurred and the duration varies, making it difficult to determine when the behavior begins and ends.Previously, researchers have tackled these challenges with various approaches. Some have utilized complex models to improve behavior classification accuracy [1, 29], while others have focused on refining temporal localization techniques [6, 21]. However, these methods often suffer from high computational costs, scalability issues, or reliance on extensive training data. These challenges under-

*These authors contributed equally to this work.
†Corresponding author

score the need for a more effective and efficient approach to temporal behavior localization in driving scenarios. In this work, we propose a novel method that addresses these challenges, which aims to improve behavior detection and localization accuracy while minimizing computational complexity, making it a promising solution for real-world applications in driver behavior analysis.

A two-stage method based on the sliding window is proposed to address the aforementioned challenges. Specifically, the approach uses the following strategies: 1) defining a fixed-duration sliding window that slides along the time axis of the video and sequentially identifies the specific behavior category within each time interval corresponding to the sliding window; 2) utilizing post connection and an adaptive thresholding scheme. The prediction results on the sliding window are determined by the information provided by three views, while the adaptive thresholding method is applied to filter out repeated false detections and make up for missed detections; 3) utilizing a candidate behavior interval splicing criterion to detect behaviors and determine their start and end times in the untrimmed video. A post connection strategy is employed to further refine the behavior boundaries from coarse localization to precise localization, thereby improving the overall performance of the model.

In summary, this paper contributes in the following aspects:

- In the natural driving action recognition task of 2024 AI City Challenge, we propose a two-stage approach based on the sliding window for the detection of distracted driving behaviors. This approach is efficient and effective, and performs well in the evaluation of 2024 AI City Challenge Track 3.
- To address the challenge of high similarity among different behaviors and interference from background classes, we introduce multi-view fusion and adaptive thresholding strategy, which effectively filters out false detections and improves missed detections.
- To address the fuzzy problem of behavior boundary positioning, we design a post-processing procedure, which further improves the overall performance of the model by refining the behavior boundary through post connection and candidate behavior splicing criteria.

2. Related Work

Our framework consists of two main components: action classification and temporal action localization. Both of these components are important research branches in computer vision, and there have been a large number of related works. In this section, we summarize the methods and related works that we were used in our research.

2.1. Action classification

Action classification is an important branch of video understanding with broad application prospects. At a macro level, mainstream action classification methods can be divided into two categories: 2D convolution-based and 3D convolution-based methods. 2D convolution-based methods extract features from each frame of the video using 2D convolutional neural networks (CNNs) and classify these features with a classifier to recognize the action in the video [2, 9, 11]. However, 2D convolution-based methods cannot consider temporal information because they can only process each frame individually, without capturing the temporal relationships between frames, which results in the model's inability to handle fast actions or changes.

3D convolution-based action recognition methods learn the temporal and spatial features of actions by performing convolution operations in both time and space dimensions, thus better capturing action information in videos. Common 3D convolution-based action recognition methods include C3D [19], Res3D [10], LTC [23], and I3D [4]. Due to the significantly larger number of parameters and computational complexity of 3D convolutional networks compared to 2D convolutional networks, some methods focus on low-rank approximation of 3D convolutional networks, such as FstCN [17], P3D [16], R(2+1)D [20], and S3D [26]. In addition, the X3D [8]method adjusts the hyperparameters of 3D CNNs to make the network more compact and efficient.

Moreover, methods based on Transformer networks, such as Video-Swin-Transformer [14] and ViViT [3], can better handle the spatio-temporal features of video sequences. In response to the characteristics of the Track3 dataset, we selected Video-swin-Transformer to avoid overfitting. [29]

2.2. Temporal Action Localization

Temporal action localization is an important task in video action recognition, which aims to determine the start and end time of a specific behavior in a video. Based on deep learning, temporal localization methods can be mainly categorized into three types: one-stage, two-stage, and multi-stage methods. One-stage network methods directly predict the start and end time of behaviors from the video sequence. This method [5] typically converts the temporal localization task into a regression problem using a similar approach to object detection. Two-stage methods typically produce temporal candidate proposals, which are subsequently classified and their temporal bounds refined. Techniques like border detection refinement [13] and sliding windows aggregation [7] have been used in earlier studies. Current investigators have put forth strategies that model action contexts using graph architectures [28] and attention mechanisms [18, 24]. Multi-stage network methods use multiple network modules to extract features and

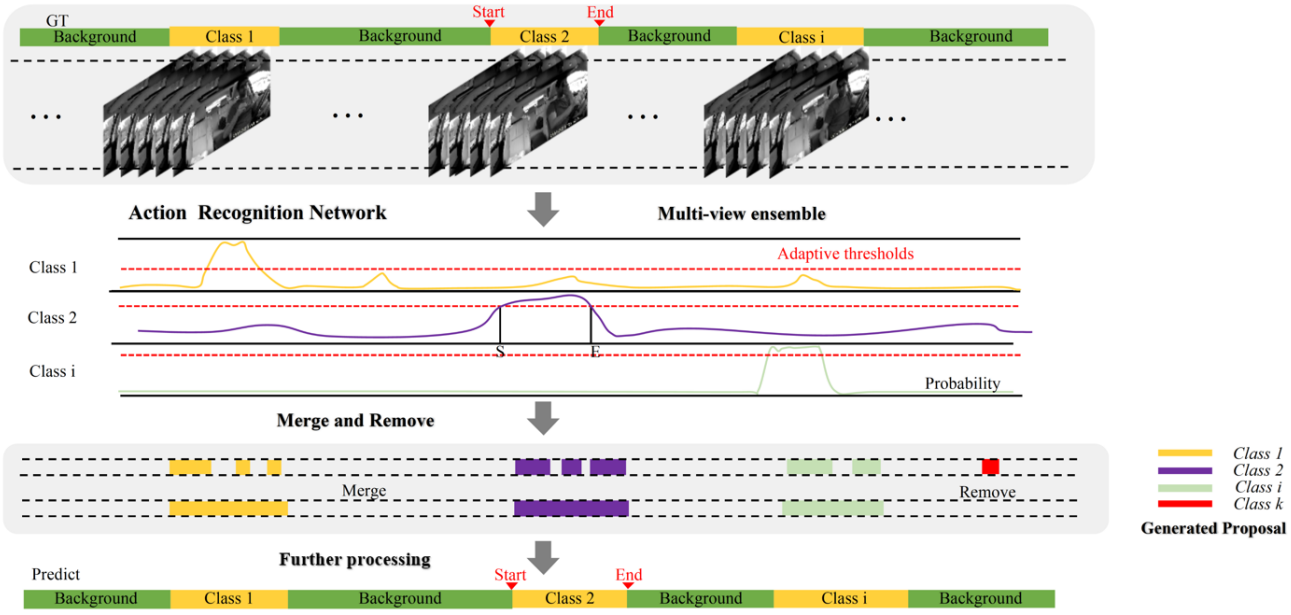

Figure 2. The framework diagram of a robust distracted action recognition method is presented. The system takes video as inputs and goes through video clips sampling , action recognition classification, multi-view fusion, adaptive thresholding filtering, merging and removing, and post connection to obtain the final results.

perform temporal localization step by step. Common multi-stage network methods include BMN [12] and G-TAD [27].

Overall, one-stage research has become increasingly mainstream, but these methods have high computational costs and require large amounts of data to train the network. Moreover, these methods cannot meet the requirements for highly accurate temporal boundaries and classification.

Therefore, we rethink the method and start with a two-stage network to simplify the temporal localization process. Compared to conventional one-stage approaches, our suggested two-stage technique has advantages.Through the division of the activity into successive stages, it lowers computing costs and facilitates more effective resource usage. It may also more accurately achieve temporal boundary localization and classification and effectively utilize less datasets.

3. Method

The natural driving behavior recognition framework proposed in this paper consists of two stages, namely the action classification stage and the temporal action localization stage. The design and improvements of each stage will be described in detail below. The input of this framework is video frames, which go through processes such as action classification, temporal action localization, and post-processing to obtain the final results.

3.1. Overall Architecture

The distraction behavior classification and temporal localization task in track3 requires accurate classification of many similar behavior categories and correct recognition and localization of distraction behavior in untrimmed videos with missing explicit boundaries or obvious object interactions. To address this challenge, a two-stage framework for distraction behavior classification and temporal localization is proposed in this paper. The framework consists of two stages: behavior classification and temporal localization with post-processing.

In the behavior classification stage, video clips are fed into the classification model to predict categories. We adopted an attention mechanism-based model as classification models. The model of attention mechanism has a strong ability to capture motion information and perform well in practice.

In the temporal localization and post-processing stage, a multi-view fusion and adaptive thresholding method is used to filter out low-confidence video clips based on the predicted classification scores. In addition, a criterion for candidate behavior interval merging and deletion is used to determine the temporal boundaries of the behavior. To further improve the overall performance of the model, a post connection strategy is employed to further improve the accuracy of behavioral boundary positioning.

3.2. Action Recognition Module

The purpose of this module is to accurately classify behaviors in video clips and provide accurate classification predictions to support subsequent temporal localization. A well-performing classifier is required for accurate temporal localization on unedited natural driving behavior videos. This section will introduce the behavior classification network used in our approach, as well as the model training strategy adopted.

Video Swin Transformer [14] is a Transformer-based 3D behavior recognition network, which not only leverages the global information modeling ability of Transformers but also employs the method of moving windows to connect across windows. This allows the model to focus on information related to adjacent windows, extending the field of perception to some extent and resulting in higher efficiency. Due to a series of advantages of Video Swin Transformer, we selected Swin-L as the backbone network for behavior recognition.

Due to the difficulty of the dataset, where different behaviors have a high degree of similarity and are interfered by the background class, the accuracy of the model under a single view is insufficient to meet the task requirements. Therefore, it is necessary to fully utilize the information from the three views. However, the difficulty of classifying the same behavior in different camera views is different. For example, it is easy to distinguish left-hand phone calls in the dashboard view, while the key item, the phone, may be occluded in the right view, making classification difficult. In order to allow the model to learn the features of different perspectives in a more targeted manner on the basis of learning all video features, we adopt a strategy of training three perspectives separately to further improve the accuracy of the model.

3.3. Temporal Localization Module

The objective of this module is to obtain the start and end times of distracted driving behaviors in untrimmed videos through temporal localization, in order to achieve more accurate results. In our framework, post-processing for temporal localization is crucial, which includes filtering out erroneous behaviors, connecting segments, and obtaining proposal behaviors. The TAL process we designed are shown in Fig. 3.

Threshold filtering. To address the issue of inaccurate predictions due to the difficulty of learning for the classifier, we propose a threshold filtering method. In general, threshold filtering is effective. However, choosing the appropriate threshold has become a tricky problem. Previous methods [22] have used the average prediction score of the classifier on all videos as the global threshold and then filtered out video segments with scores below the average threshold. We call this solution Method1, and its formula is as follows:

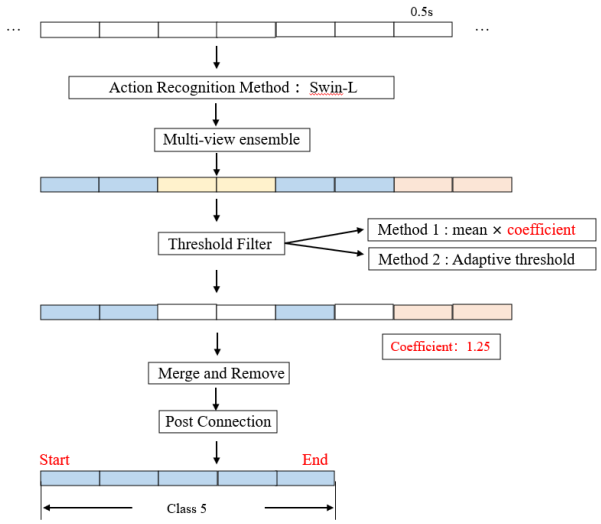

Figure 3. The temporal localization and post-processing module, which includes multi-view ensemble, threshold filtering, clip merging and removing, post connection.

$$Thr = \frac{\sum_{c=0}^{N=16} \sum_{j=0}^{M} \max_{i=1,2,3} \left\{ p_{ij}^{c} \right\}}{N \times M} \tag{1}$$

where N represents the total number of categories and M represents the total number of video clips per clip. p_{ij}^{c} represents the classification prediction score of the j-th clip in the i-th perspective for the c-th category.

However, due to differences in the duration and classification difficulty of different behaviors, some video segments have low confidence but are classified correctly, while some video segments have high confidence but are classified incorrectly. Therefore, we propose an adaptive thresholding approach. We take the top-M scores for each class according to the time axis and filter out video segments with scores lower than the class-specific threshold. We call this solution Method2, and its formula is as follows:

$$Thr_c = \mathbf{Sort}\left(\max_{i=1,2,3} \left\{ p_{ij}^{c} \right\}, \boldsymbol{M} \right) \tag{2}$$

where $Sort(\cdot, M)$ means to sort all the predicted scores of the c-th category in descending order and take the M-th value. p_{ij}^{c} represents the classification prediction score of the j-th clip in the i-th perspective for the c-th category.

By using this method, errors in video segments can be filtered more accurately, resulting in more precise start and end times for distracted behaviors. This can help improve the performance and effectiveness of our model.

Merging and Removing. The objective of this module is to concatenate short video segments into longer candidate segments. Although threshold filtering can filter out a large

number of misclassified and redundant video segments, the filtered segments of each class are not contiguous in time, making it difficult to determine the temporal boundaries of behaviors. Therefore, it is necessary to splice discrete video segments according to reasonable rules. To this end, we have designed splicing criteria: if the time interval dt1 between two segments of the same category is less than T1, these two segments are spliced together. Through this splicing operation, the extraction of candidate segments is completed. In addition, since distracted behaviors generally last no less than 3 seconds, we delete candidate segments with a duration less than 3 seconds. In practice, we found that after connecting the segments, there may be overlapping predicted results. To solve this problem, we have designed a simple filtering criterion inspired by the NMS [15], which is to retain the first occurring behavior and remove the later occurring behavior in the overlapping behaviors. Finally, we successfully obtained candidate results for each category and improved the accuracy of the algorithm.

Further processing. Taking into account the influence of actual labels and datasets, this module is designed to further refine the behavior boundaries from coarse localization to fine localization to improve the final recognition accuracy. Given the irregular intervals at which actual behavioral segments occur, we perform a secondary connection. Previously, clips are connected to short proposals. Now, for these short proposals, we conduct the post connection. If the interval between two short proposals is less than dt2, we connect the two short proposals, until there is no short proposals to connect. The empirical value of dt2 is 20s. Now, the final proposals extraction is complete.

4. Experiment

In this section, we will present in detail the threshold selection strategy, the separate training strategy and the post connection strategy that we proposed, and explain some experimental details. Subsequently, we will demonstrate the effectiveness of the proposed method.

4.1. The Dataset of AI City Track 3

The dataset used in our study consists of a total of 594 videos, which were captured from 99 different drivers. These video clips total about 90 hours. Each driver performed 16 different tasks, such as talking on the phone, eating, and reaching back, once in a random order. The data collection was done using three cameras mounted in the car, recording from different angles in synchronization. Specifically, the videos in the dataset have a resolution of 1920×1080 pixels, a frame rate of 30 frames per second, and are divided into three categories: A1 for training and A2/B for testing. The training set contains information about the start and end times of each behavior, as well as the category of the behavior, as shown in Tab. 1. This dataset provides a comprehensive and diverse benchmark for evaluating our proposed methods.

ID	Description	ID	Description
0	Normal Driving	8	Adjust control panel
1	Drinking	9	Pick up from floor(Dri)
2	Phone Call(Right)	10	Pick up from floor(Pas)
3	Phone Call(Left)	11	Talk to Pas(right)
4	Eating	12	Talk to Pas(backseat)
5	Text(Right)	13	yawning
6	Text(Left)	14	Hand on head
7	Reaching behind	15	Sing or dance with music

Table 1. 16 distracted actions. Label 0 is not considered for the evaluation.

4.2. Experiment Setting

During the data preprocessing stage, the videos were sampled at a rate of 8 frames per video. During training, we adopt a pretrain-and-finetune manner, that is, putting the data from three perspectives together for pre training, and then fine tuning them separately. Training three perspectives separately enhances the model's ability to learn distinct features from each perspective,thereby further improving the accuracy of the model. For the pre training process, we use a pre trained Video Swin Transformer model on Kinetics-400 [4]. All training and inference were performed on 4 NVIDIA Tesla T4 GPUs, with each GPU having 15GB of memory.

During training, we use 2 second as a clip, which contains 16 frames, and the temporal order of clips is randomly shuffled when inputting them into the classification network. This can prevent the network from overfitting to a certain class. During inference, we use 0.5 seconds as a clip and perform inference in chronological order.

4.3. Metrics

Evaluation for track 3 is based on model activity identification performance, measured by the average activity overlap score, which is defined as follows. Given a ground-truth activity g with start time g_s and end time g_e, we will find its closest predicted activity match as that predicted activity p of the same class as g and highest overlap score os , with the added condition that start time p_s and end time p_e are in the range $[g_s\text{-}10s, g_s\text{+}10s]$ and $[g_e\text{-}10s, g_e\text{+}10s]$, respectively. The overlap between g and p is defined as the ratio between the time intersection and the time union of the two activities, i.e. [25]. The Intersection over Union(IoU)'s formulation is as follows:

$$os(p,g) = \frac{\max(\min(ge,pe)\text{-}\max(gs,ps),0)}{\max(ge,pe)\text{-}\min(gs,ps)} \quad (3)$$

Swin [14]	Respective Train	Adaptive Thresholding	Post Connection	Scoce
✓				0.4755
✓	✓			0.5066
✓	✓	✓		0.6184
✓	✓	✓	✓	0.6272

Table 2. Ablation on the Verification set with different correction strategies.

The Average Intersection over Union (AIOU) is the average IOU calculated for each video.

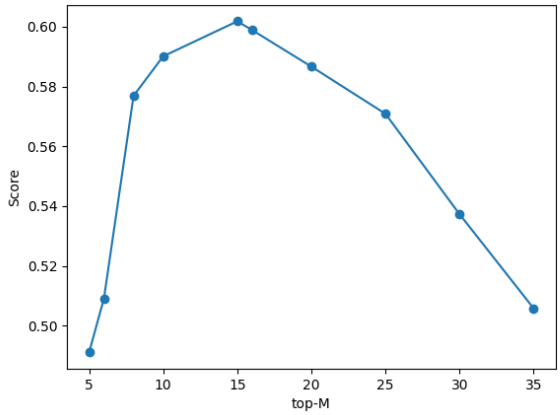

Figure 4. Results of selecting top-M on the validation set.

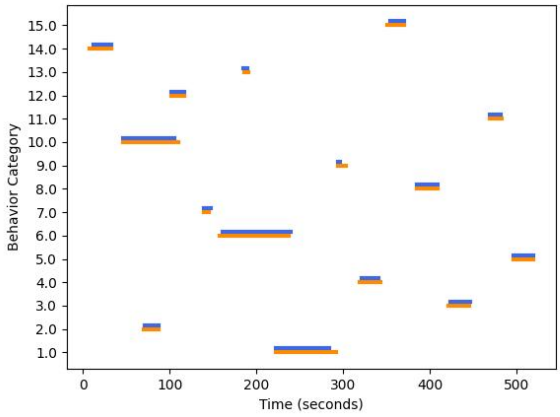

Figure 5. Method results visualization on one of the verification data set.The orange lines are the ground-truth distribution, and the blue lines show the final predict results.

4.4. Main results

In order to better simulate the data distribution of the test set A2 and verify the effectiveness of the model, we selected all videos of A1_7 set as the verification set. The ablation on the verification set are shown as Tab. 2. Obviously, training the data set respectively in three perspectives can enhance the recognition module's functionality. Post connection module improves the model to some extent. Adaptive thresholding strategy is quite effective, significantly improves the Score from 0.5066 to 0.6184. Additionally, Fig. 4 indicates that the adaptive thresholding is highly sensitive to the top-M hyperparameter, as the selection of the threshold is crucial in locating the regions where the behavior occurs in the sliding window temporal localization algorithm. When M is too small, the threshold is too high, resulting in decreased recall due to shorter behavior regions. When M is too large, the threshold is too low, resulting in decreased precision due to longer behavior regions. There is an optimal M value that allows the calculated threshold to optimize the framework's performance. Based on the results of the ablation experiments, we selected the hyperparameters that produced the best performance.

To more intuitively show the effect of our model, the performance of the model is visualized in Fig. 5. It's noticeable that our method performs excellently in classification precision. This is due to the fact that we fully utilize the outcomes of action recognition from three perspectives and employ numerous useful techniques. The adaptive thresholding approach, in particular, significantly enhances the model.

4.5. Final Ranking

The performance of our framework and models has been gradually improved through continuous optimization, making it more practical in real-world scenarios. The specific Test A results can be seen in Tab. 3, which lists the scores of the top 10 teams. We finally achieve a score of 0.6080.

5. Conclusion

The detection and recognition of distracted driving behaviors has emerged as a new vision task with the rapid development of computer vision. Our proposed adaptive

Rank	Team	Score
1	TeleAI	0.8282
2	supermonkey	0.8213
3	yptang	0.8149
4	Rockets	0.8045
5	SKKU-AutoLab	0.7798
6	Bumblebee_AIO	0.7624
7	boat	0.6844
8	**MCPRL(ours)**	**0.6080**
9	zzl	0.5963
10	USTC-IAT-United	0.2307

Table 3. Top 10 Leaderboard of AI City 2024 Track 3 Naturalistic Driving Action Recognition.

distracted driving behavior recognition system achieved good performance in the AICITY2024 Task3 competition. This task is a temporal action localization problem, and for untrimmed naturalistic driving videos, we adopted a two-stage framework for distracted driving behavior classification-temporal localization based on sliding window algorithm, which is concise and efficient. To address the problem of high similarity between different behaviors and background class interference, we used multi-view fusion and adaptive thresholding methods, effectively filtering out false positives and missed detections. In addition, to solve the problem of blurred behavior boundary positioning, we designed a set of post-processing procedures, including post connection and candidate behavior splicing criteria, achieving the goal of coarse-to-fine localization of behavior boundaries. Particularly, we conducted detailed ablation experiments to validate the effectiveness of the improved methods. In the future, we will continue to focus on balancing the accuracy and efficiency of temporal localization, in order to apply the methods proposed in this paper to practical applications.

6. Acknowledgements

This work is supported by Chinese National Natural Science Foundation under Grants 62076033.

References

[1] Erkut Akdag, Zeqi Zhu, Egor Bondarev, and Peter H. N. de With. Transformer-based fusion of 2d-pose and spatio-temporal embeddings for distracted driver action recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5453–5462, 2023. 1

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 2

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 5

[5] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1130–1139, 2018. 2

[6] Xiaodong Dong, Ruijie Zhao, Hao Sun, Dong Wu, Jin Wang, Xuyang Zhou, Jiang Liu, Shun Cui, and Zhongjiang He. Multi-attention transformer for naturalistic driving action recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5435–5441, 2023. 1

[7] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 768–784. Springer, 2016.

[8] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 2

[9] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 971–980, 2017. 2

[10] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017. 2

[11] Dongliang He, Zhichao Zhou, Chuang Gan, Fu Li, Xiao Liu, Yandong Li, Limin Wang, and Shilei Wen. Stnet: Local and global spatial-temporal modeling for action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8401–8408, 2019. 2

[12] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898, 2019. 3

[13] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *European Conference on Computer Vision*, 2018. 2

[14] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 2, 4, 6

[15] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, volume 3, pages 850–855. IEEE, 2006. 5

[16] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 2

[17] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4597–4605, 2015. 2

[18] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13526–13535, 2021. 2

[19] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[20] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[21] Manh Tung Tran, Minh Quan Vu, Ngoc Duong Hoang, and Khac-Hoai Nam Bui. An effective temporal localization method with multi-view 3d action recognition for untrimmed naturalistic driving videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3167–3172, 2022. 1

[22] Manh Tung Tran, Minh Quan Vu, Ngoc Duong Hoang, and Khac-Hoai Nam Bui. An effective temporal localization method with multi-view 3d action recognition for untrimmed naturalistic driving videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3168–3173, 2022. 4

[23] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517, 2017. 2

[24] Lining Wang, Haosen Yang, Wenhao Wu, Hongxun Yao, and Hujie Huang. Temporal action proposal generation with transformers. *arXiv preprint arXiv:2105.12043*, 2021. 2

[25] Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Pranamesh Chakraborty, Sanjita Prajapati, Quan Kong, Norimasa Kobori, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Ganzorig Batnasan, Fady Alnajjar, Ping-Yang Chen, Jun-Wei Hsieh, Xunlei Wu, Sameer Satish Pusegaonkar, Yizhou Wang, Sujit Biswas, and Rama Chellappa. The 8th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024. 1, 5

[26] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 2

[27] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10156–10165, 2020. 3

[28] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7094–7103, 2019. 2

[29] Hangyue Zhao, Yuchao Xiao, and Yanyun Zhao. Pand: Precise action recognition on naturalistic driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3291–3299, 2022. 1, 2