MV-TAL: Multi-view Temporal Action Localization in Naturalistic Driving

Wei Li¹* Shinmin Chen¹* Jianyang Gu¹,²* Ning Wang¹,³ Chen Chen¹ Yandong Guo¹
¹OPPO Research Institute. ²Zhejiang University. ³East China University of Science and Technology
{liweil9, chenshimin1, chenchen, guoyandong}@oppo.com
gujianyang@zju.edu.cn wangning12@mail.ecust.edu.cn

Abstract

Human risky behavior in driving is an important visual recognition problem. In this paper, we propose a multi-view temporal action localization system based on the grayscale video to achieve action recognition in naturalistic driving. Specifically, we adopted SwinTransformer as feature extractor, and a single framework to detect boundary and class at the same time. Also, we improve multiple loss function for explicit constraints of embedded feature distributions. Our proposed framework achieves the overall F1-score of 0.3154 on A2 dataset.

1. Introduction

With the development of automation, computer vision technologies has achieved great progresses on several tasks related to the general vehicle structures, including vehicle classification [7, 15, 48], detection [16, 24], tracking [29, 39], trajectory prediction [3, 38] and fine-grained re-identification [25, 26]. However, driver distracted behavior detection, which takes place inside vehicles and plays an essential role in human-vehicle communication, dynamic driving adaptation and safety, is still understudied.

Driver behavior recognition is closely linked to the broader field of action recognition, where the performance numbers have rapidly increased due to the rise of deep learning. Such models are data-hungry and are often evaluated on large, color-based datasets with a carefully selected set of highly discriminate actions, usually originated from Youtube such as ActivityNet-1.3 [2] and HACS [50]. To locate the spatial positions and temporal boundaries of each action in untrimmed videos is a challenging task. And there is still a lot of room for the research on driver activity understanding. In the Driver Action Recognition field, different from traditional action recognition tasks, it involves timely safety issues which make it very sensitive to action boundary, so the track 3 [32] requires the recognition error to be within 1s. What’s more, it is difficult to recognize all actions in a single camera view due to occlusion. To enhance recognition, some datasets [28] for autonomous driving action recognition propose multi-view recognition to increase action diversity.

In this paper, we constructed a MV-TAL (multi-view temporal action localization) system based on the grayscale videos inside the car. The framework of our MV-TAL system is shown in Figure 1. We construct it with feature sequences extracted from raw video by SwinTransformer [27] classifiers in different views and clip lengths. Then a temporal action localization algorithm is applied to detect the action boundaries and classes at the same time. Multiple metric learning loss functions are introduced to explicitly optimize the embedded feature distributions. Last, we ensemble the results of different views and temporal action detection models which complement each other.

To sum up, our contribution are as follows:

1. To improve feature representation, we construct 12 diverse features which can complement each other.
2. In order to maximize the value of features from different views, we propose to improve the loss function for explicit constraints of embedded feature distributions.
3. We take full advantage of features with different views in a single network, which simplifies computation cost and achieves great performance.
4. We employ different clip duration features as auxiliary features, which enable the model well complemented in localization and classification performance.

2. Related Work

2.1. Action Recognition

Traditional video-based action recognition consists of action classification [6, 9, 22, 31, 35, 40, 41, 43], 3D-skeleton action classification [20, 47], temporal action localization [13, 14, 36, 45, 49], and spatio-temporal action localization [8, 17–19, 21, 33, 44]. For driver temporal action detec-
3. Method

3.1. Feature Engineering

In order to construct diverse features, we used classification models to learn different action characteristics. In this section, considering that different views and clip lengths may have different representations. First, according to our own observations, we mask some actions from a specific view. For example, The “Text” which is indistinguishable from the dashboard view will be masked. The mask information are shown in table 2, and the actions which are not masked in each view are not shown in it. Then, we observe that start, end, and middle representations of some labeled actions like “Eat” are quite different. As a result, we designed view-wise classification with 3 views, 2 clip lengths (3 and 6 seconds) and 2 label definitions, leading to 12 diverse feature extractors, and sample unlabeled segments to be negative clips. The first label definition refers to the origin label. In second label definition, we split an action into start, middle and end segments so that features can help to localize action boundaries. We set different strides for positive and negative to overcome the shortness of data imbalance. The data acquisition details are shown in table 1.

Considering the effectiveness and diversity, the distinguishing view classification model with 3-second and 6-second video clips are adopted to extract features with stride 32. It is noticed that we just use the model of the corresponding view to extract the features in distinguishing view classification, so that each user id with occlusion status will have 12 features consisting of 3 views and 4 kind of clips.

3.2. Temporal Action Detection

3.2.1 Temporal Proposal Generation

Inspired by [5], we adopt a similar framework to generate temporal proposals and semantic labels in a unified network.

As shown in 1, we propose a MV-TAL (multi-view temporal action localization) model to detect actions inside the car. Following Faster-TAD [5], we adopt Confidence-Matching mechanism [23] to generate proposals. Proposal Generation Mechanism contains two branches, Temporal Evaluation Module and Proposal Evaluation Module. The Proposal Generation Mechanism generates coarse proposals, which are then refined by Proposal Refinement Module.
Table 1. The stride and class number information of six methods.

<table>
<thead>
<tr>
<th>method</th>
<th>view split</th>
<th>start end split</th>
<th>clips</th>
<th>stride</th>
<th>class number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>pos-start</td>
<td>pos-mid</td>
</tr>
<tr>
<td>1</td>
<td>×</td>
<td>×</td>
<td>3-second</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>×</td>
<td>×</td>
<td>6-second</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>✓</td>
<td>×</td>
<td>3-second</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>✓</td>
<td>×</td>
<td>6-second</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>5</td>
<td>✓</td>
<td>✓</td>
<td>3-second</td>
<td>0.4</td>
<td>0.25</td>
</tr>
<tr>
<td>6</td>
<td>✓</td>
<td>✓</td>
<td>6-second</td>
<td>0.2</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Figure 2. Different view mask in action “Adjust control panel” (top) and “text (left hand)” (bottom). The figure with red bbox means the action in this view are difficult to recognize, while the green bbox has the opposite meaning.

Table 2. The mask information of each view, × means mask and ✓ means no-mask.

<table>
<thead>
<tr>
<th>Index</th>
<th>Name</th>
<th>Dash</th>
<th>Rear view</th>
<th>Right win</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Normal Forward Driving</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>5</td>
<td>Text (Right)</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>6</td>
<td>Text (Left)</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>9</td>
<td>Adjust control panel</td>
<td>×</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>10</td>
<td>Pick up from floor (D.)</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>11</td>
<td>Pick up from floor (P.)</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>16</td>
<td>Singing with music</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
</tr>
</tbody>
</table>

Evaluation Module (TEM) and Proposal Evaluation Module (PEM). Temporal Evaluation Module aims to evaluate the starting and ending probabilities for all temporal locations in untrimmed video. In Proposal Evaluation Module, we adopt SGAlign [30] Block to generate Boundary-Matching (BM) confidence map, which aims to evaluate the probability of proposal globally. We use boundary probability sequences and BM confidence map to generate proposals during post processing. For proposal Classification, we adopt Context-Adaptive Proposal Module [5] to encode proposal features. It should be noted that since there is little related information related to the background proposal adjacent to the positive proposals in this task, we did not utilize Proximity-Category Proposals Block. For Proposal Regression Refinement, we adopt Local-Global Temporal Encoder [30] to model video feature sequence locally and globally. Then, we further employ Temporal Boundary Regressor Block [30] to refine coarse proposals.

In order to detect actions inside the car with multi-views videos, we bring three improvements. First, we utilize three view videos as inputs to generate proposals and semantic labels. In this way, model can fuse multi-view information and learn better results. Secondly, we adopt three Base Module and three Local-Global Temporal Encoder to sep-
arately encode different view features. This mechanism al-
low model firstly learn the differences of features from dif-
ferent views inputs, and then learn fusing features to get
better results. Last but not least, we employ 3s features as
auxiliary features. By fusing features of different clip du-
ration, the model is well complemented in localization and
classification performance.

3.2.2 Proposal Classification

Classification is also an essential part in the temporal ac-
tion detection process. Different from common temporal ac-
tion detection datasets, where each action category contains
sufficient samples, Track 3 only provides 30 ground truths
for each category. Besides, due to the influence of camera
poses, samples of different categories under the same view
share more similar appearances, compared with those of the
same category under different views. The above factors pre-
vent the classification model from getting clear classifica-
tion results and classification results of 4 methods.

Table 3. The results of six methods.

<table>
<thead>
<tr>
<th>method</th>
<th>view</th>
<th>top1_acc</th>
<th>top5_acc</th>
<th>mean_acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>3s &amp; origin label &amp; cat view</td>
<td>reearview</td>
<td>22.93</td>
<td>49.42</td>
<td>8.52</td>
</tr>
<tr>
<td>3s &amp; origin label &amp; split view</td>
<td>window</td>
<td>44.11</td>
<td>73.63</td>
<td>44.06</td>
</tr>
<tr>
<td>6s &amp; origin label &amp; split view</td>
<td>dashboard</td>
<td>52.19</td>
<td>88.92</td>
<td>40.12</td>
</tr>
<tr>
<td>3s &amp; split label &amp; split view</td>
<td>reearview</td>
<td>64.74</td>
<td>88.12</td>
<td>48.15</td>
</tr>
<tr>
<td>3s &amp; split label &amp; split view</td>
<td>window</td>
<td>52.03</td>
<td>79.98</td>
<td>51.83</td>
</tr>
<tr>
<td>6s &amp; split label &amp; split view</td>
<td>dashboard</td>
<td>50.22</td>
<td>85.09</td>
<td>48.80</td>
</tr>
</tbody>
</table>

In the actual calculation process, the weighting factors are
assigned as $\alpha_p = [1 + m - s^p_n]_+$ and $\alpha_n = [s^l_s + m]_+$. The margins are set as $\Delta_p = 1 - m$ and $\Delta_n = m$. The above loss functions are grouped in multiple ways to pro-
duce different TAD models. We employ model ensemble to
aggregate the advantages of one another.

3.3. Ensemble

In the Feature Engineering mentioned in Chapter 3.1, we
can not only generate discriminative features for temporal ac-
tion detection, but also get the classification results cor-
responding to each feature. In this section, for the different
method mentioned in Table 1, we synthesize the proposal
classification results in Chapter 3.2.2 and the classification
results in the classifier to form the final classification results,
and apply soft-NMS [1] to the proposal localization results
with different thresholds for different category. Besides, to
increase model diversity and maximize the value of features
from different views, we also ensemble the proposal local-
ization results and classification results of 4 methods.

4. Experiment

4.1. Classifier

In this section, we present the results of 12 classifier men-
tioned in 3.1, as shown in Table 3. It shows that
the model with different view can get better performance.
Specifically, we use the user_id 35133 as test dataset and
others as train dataset from A1 dataset, and we set the inter-
val=8 in 3-second clips training and interval=4 in 6-second
clips training. We run all experiments on a machine with 8
NVIDIA GTX1080Ti GPU.
Table 4. The results of ensemble, “cat” means we cat features from different view as training features, and & means using the two stream input of two kind of features. “-se” means use the split label with start and end and others means not.

<table>
<thead>
<tr>
<th>Features</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>3s-se 6s 3s&amp;6s cat 3s cat 6s cat 3s &amp; 6s</td>
<td>0.2346</td>
<td>0.4375</td>
<td>0.3055</td>
</tr>
<tr>
<td>✓ ✓ ✓ ✓ ✓ ×</td>
<td>0.2346</td>
<td>0.4330</td>
<td>0.3043</td>
</tr>
<tr>
<td>✓ ✓ ✓ ✓ ✓ ×</td>
<td>0.2291</td>
<td>0.5062</td>
<td>0.3154</td>
</tr>
</tbody>
</table>

4.2. Ensemble

In this section, we present the best ensemble result of different temporal action detection models trained with different features, which shows the multiple model using different features can complement each other, as shown in Table 4.

5. Conclusion

In this paper, we present our approach for the CVPR2022 Workshop AICity Challenge Track 3. A driver temporal action detection system is proposed for naturalistic driving action recognition. We propose MV-TAL network to detect temporal actions with multi-views. Different clip duration features are employed as auxiliary features, which enable the model well complemented in localization and classification performance. What’s more, we propose to involve metric learning loss functions for explicit constraints of embedded feature distributions. Also, we construct multi features to improve diversity of feature representation. Our network can aggregate features with different information and further improve the performance. Our strategies have shown great performance in classification and localization.

References


Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1


Mohammed Shaiqur Rahman, Archana Venkatachalapathy, Anuj Sharma, Jiyan 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, 2022. 1


