Real-Time Driver Monitoring Systems through Modality and View Analysis

Yiming Ma  
University of Warwick  
Coventry, UK

Victor Sanchez  
University of Warwick  
Coventry, UK

Soodeh Nikan  
Ford Motor Company  
USA

Devesh Upadhyay  
Ford Motor Company  
USA

Bhushan Atote  
University of Warwick  
Coventry, UK

Tanaya Guha  
University of Glasgow  
Glasgow, UK

October 19, 2022

Abstract

Driver distractions are known to be the dominant cause of road accidents. While monitoring systems can detect non-driving-related activities and facilitate reducing the risks, they must be accurate and efficient to be applicable. Unfortunately, state-of-the-art methods prioritize accuracy while ignoring latency because they leverage cross-view and multimodal videos in which consecutive frames are highly similar. Thus, in this paper, we pursue time-effective detection models by neglecting the temporal relation between video frames and investigate the importance of each sensing modality in detecting drivers’ activities. Experiments demonstrate that 1) our proposed algorithms are real-time and can achieve similar performances (97.5% AUC-PR) with significantly reduced computation compared with video-based models; 2) the top view with the infrared channel is more informative than any other single modality. Furthermore, we enhance the DAD dataset by manually annotating its test set to enable multiclassification. We also thoroughly analyze the influence of visual sensor types and their placements on the prediction of each class. The code and the new labels will be released.

1 Introduction

While the proliferation of on-road traffic has dramatically benefited society, it has also increased fatal traffic accidents. According to the World Health Organization, until June 2022, there have been about 1.3 million casualties and more than 20 million injuries from car crashes every year, and these accidents cost most countries approximately 3% of their GDPs. In these crashes, one of the dominant contributing factors is human errors from inattention, such as preoccupation with mobile phones while driving. Hence, for L2+ self-driving-enabled cars, it is crucial to develop effective driver monitoring systems (DMSs) to estimate the drivers’ readiness for driving and take over the control when necessary to prevent accidents.

As an essential information source, vision is often exploited by DMSs to detect drivers’ non-driving-related activities (NDRAs). DAD [1] is one of the latest video databases for vision-based monitoring systems. In addition to the rich diversity of NDRAs in the training set, its test set also contains unseen types of actions. Since there can be unboundedly many actions that drivers may conduct and the unseen cases in the test set of DAD enable it to better generalize to realistic driving, we decide to establish our work on this open-set recognition dataset.
Table 1: Details of the non-driving-related activities (NDRAs) in the DAD dataset [1]. Activities in black are present in both the training and test sets, and those in red are exclusive to the test set. The original DAD dataset only provides binary (normal/anomalous) labels for the test set as it is initially developed for anomaly detection.

However, the test-set labels of DAD are binary, only indicating whether drivers are participating in NDRAs or not. This inadequacy hinders the multiclassification of drivers’ activities, which is essential since different activities may not be equally hazardous. Hence, we manually annotate the test set with the specific task names to allow the recognition of these actions and class-based evaluation of DMSs.

On the other hand, the latest vision-based DMSs are not sufficiently efficient. In-cockpit vision systems (including DAD) usually consist of visual sensors of various types (e.g. RGB) installed at different locations (e.g. top) to provide videos of the driver from diverse streams and views. Thus, to maximize the accuracy of activity detection, existing approaches often employ spatial and temporal information from all modalities and views. These methods unavoidably introduce billions of floating-point operations (FLOPs) into inference, which embedded devices cannot perform in real time. Therefore, in this paper, we aim to develop a real-time DMS by neglecting the temporal dimension and only employing the most informative vision source. Experiments on DAD demonstrate that 1) neighboring frames within a sequence are highly resembling; 2) single-modal architectures based on the top view and the infrared modality can also achieve state-of-the-art performances. These two findings validate our approach.

Our contributions are summarized as follows:

1. We propose efficient image-based DMSs with comparable performances (97.5% AUC-PR & 95.6% AUC-ROC) with state-of-the-art video-based models. Unlike other methods’ substantial computation load, our models’ low latency makes them deployable in the real world.

2. We analyze the performance of our models on each view and modality and provide the most economical solution to the placement (top) and the type (IR) of cameras that DMSs should leverage.

3. We annotate the test set of DAD with specific activities, thereby enabling the multiclassification of drivers’ actions on it. This particularized labelling allows detailed analysis of algorithms based on each class, and detecting the most dangerous distractions (e.g. using mobile phones) can be prioritized.

2 Related Work

2.1 In-Car Vision-Based Datasets

Early datasets focus only on parts of the driver’s body, like the head [2]–[6] or hands [7]–[9]. Although these may have contributory values in other tasks, such as gesture recognition and pose study, only a narrow range of the whole-body movements are typically captured, thereby covering very few NDRAs.

Later datasets that followed [10]–[12] instead concentrate on body actions. AUC-DD [10] is based on a single sensing modality, with images collected from a side view perspective and the RGB stream. By contrast, DMD [12] is a multimodal and video-based dataset, currently the largest for SAE L2-L3 autonomous driving. It consists of 41 hours of videos recorded from three different
views and three distinct channels. However, its training set and test set share the same classes of activities, so models trained on it may be overfitted to these seen actions and thus fail to generalize well in real-world scenarios. In comparison, DAD [1] is devised for open-set recognition. In addition to the categories in the training set, its test set also comprises unseen types of activities, as illustrated in Table 1. Therefore, this database is more appropriate for estimating DMSs’ real-world performances. However, the DAD test labels only indicate whether or not a driver is engaging in driving or NDRAs, without specifying the categories of actions. This binary labelling confines its usage to anomaly detection. As a result, DMSs developed using this dataset remain oblivious to the type of potentially distracting activity. This granularity may be critical in designing a composite attention metric where the different types of NDRAs may have different sensitivities. Hence, we found that a richer set of labels is essential for studying NDRAs and their impacts on overall driver attention.

2.2 Vision-Based Driver Monitoring Systems

Image-based DMSs. [10] proposes an approach based on object detection, in which the hand and face areas of the driver in the image are detected first and then fed into an ensemble of convolutional neural networks (CNNs). Established on this architecture, a model proposed in [13] further includes a skin segmentation branch to resolve the issue of variable lighting conditions that imposes performance degradation onto RGB-based models. Contrary to multi-branch structures, [14] leverages classic CNN classifiers such as VGG [15], also leading to state-of-the-art results. This finding motivates us to establish our work on pre-trained ResNets [16] and MobileNet [17].

Video-based DMSs. Currently, most video-based methods [1], [12] leverage 3D CNN classifiers pre-trained on Kinetics-600 [18], [19]. [12] proposes a DMS on the DMD dataset [12] that first exploits 3D CNNs (MobileNet-V2 [17] and ShuffleNet-V2 [20]) as the backbone to extract spatial-temporal features and then uses a consensus module to further capture the temporal correlations between frames. 3D ResNet-18 [16], 3D MobileNet-V2 [17] and 3D ShuffleNet-V2 [20] are utilised in [1] to extract embeddings and contrastive learning with noise estimation [21] is adopted to optimize the cosine similarities of these representations. However, [22] suggests that spatial details are more influential than temporal relations. We also prove that high-level similarities exist between consecutive frames. These two observations indicate that leveraging all frames within a video clip may not be cost-effective, so an image-based DMS does not need to make inferences for each frame, thereby showing the potential of real-time performances.

2.3 Multimodal Feature Fusion

The fundamental problem in multimodality is fusion. Previous DMSs with multimodality [1], [12] are based on decision-level fusion, while combing multimodal features is rarely studied. We regard multimodal feature fusion as a particular case of general feature fusion, and various approaches have been proposed during its evolution. Earlier ones leverage linear operations such as concatenation (e.g. Inception [23]–[25]) and addition (e.g. ResNet [16], FPN [26], U-Net [27]). This rigid linear aggrega-
Table 2: Examples of the new labels of the test set of DAD [1]. The original version does not particularize the types of NDRAs, so we review each video clip to label it. This more detailed version allows multiclassification on DAD and class-based evaluation.

| Video ID | Frames       | Original Labels | New Labels          |
|----------|--------------|-----------------|---------------------|
| rec1     | 0 - 164      | Anomalous       | Adjusting Radio     |
| rec1     | 165 - 513    | Normal          | Normal Driving      |
| rec1     | 514 - 1150   | Anomalous       | Drinking            |
| rec1     | 1151 - 1831  | Anomalous       | Normal Driving      |
| rec1     | 1832 - 2336  | Anomalous       | Adjusting side mirror |
| rec1     | 2337 - 2886  | Anomalous       | Normal Driving      |
| rec1     | 2887 - 3688  | Anomalous       | Reading             |
| rec1     | 3689 - 4399  | Anomalous       | Normal Driving      |

Table 3: Labels utilized during training and testing. We transform the original binary classification problem into a multiclassification one by annotating the test set. For convenience, we combine some of them, so there are seven classes (one for driving attentively and six for NDRAs) in the training set and one additional named “unseen” in the test set to accommodate any new NDRAs.

| Activities in Training Set          | Activities in Test Set | NDR? | Notation |
|-------------------------------------|------------------------|------|----------|
| Normal driving                      | Normal driving         | ✓    | c₁       |
| Talking on the phone                | Talking on the phone   | ✓    | c₂       |
| Messaging                           | Messaging              | ✓    | c₃       |
| Talking with passengers             | Talking with passengers | ✓   | c₄       |
| Reaching behind                     | Reaching behind        | ✓    | c₅       |
| Adjusting radio                     | Adjusting radio        | ✓    | c₆       |
| Drinking                            | Drinking               | ✓    | c₇       |
| Unseen                              |                        | ✓    | c₈       |

3 Methodology

3.1 Data Labelling

Annotating the test set. The original test-set labels of DAD [1] take only two values – “anomalous” and “normal”, denoting whether or not the driver is engaging in NDRAs. We manually scrutinize each video clip (from both views and both modalities) and label it with the corresponding class (see Table 1). During this process, we find that some activities, such as adjusting the radio, can only be confirmed from the top view, while other tasks, like talking to passengers, are only observable from the front view. This discovery demonstrates that even for human annotators, both views are required to determine the driver’s state, so cross-view models have the potential to perform better. We also find that in addition to those NDRAs listed in Table 1, there are two new tasks – “looking for something” and “yawning”, so in fact, there exist 26 different categories of NDRAs in the test set. Table 2 shows a slice of our annotations.

Labels in training and testing. In the test set, there are video clips in which drivers suddenly switch to the other hand to hold their phones. Therefore, we neglect the effect of different hand involvement to allow training and testing models effortlessly. To be more specific, we merge “messaging left” and “messaging right” into “messaging” and “talking on the phone - left” and “talking on the phone - right” into “talking on the phone”. As for unseen NDRAs, we combine all of them into a new class, “unseen”. The convention is illustrated in Table 3.

3.2 Model Structures

Assuming $X \in \mathbb{R}^{C \times T \times H \times W}$ is a short video clip, where $T$ is the number of frames, $C$ indicates the number of channels, and $H$ and $W$ are the height and width, respectively. Our image-based models neglect the temporal dimension and consider only its last frame, i.e., $X_{:, -1, :}$, which degenerates into an image. Following previous works [1], [12], [14], we leverage pre-trained convolutional classifiers as the base encoder $E$ and fully connected neural networks (FCNs) as the predication layers $P$.

The unimodal case. Given a single modality image $X$,
we exploit the base encoder to extract features
\[ f = \mathcal{E}(X). \]

Then, the multi-scale channel attention module (MS-CAM) proposed in [30] is leveraged to introduce global and local attention on \( f \). Based on the squeeze-and-excitation block [28], it adds one more branch without the global pooling to preserve local information. Features from these two branches are fused via addition and then fed into the sigmoid function to generate weights \( W \). The features with attention \( f' \) are calculated by
\[ f' = f \otimes W. \]

where \( \otimes \) stands for the element-wise multiplication. At last, \( f' \) is passed to \( \mathcal{P} \) to make the inference \( y = \mathcal{P}(f') \in \mathbb{R}^7 \). The whole process is illustrated in Fig. 2.

The multimodal scenario. We utilise the cross-view and cross-modality case to illustrate our multimodal approach. Suppose \( X^{T, IR}, X^{T, D}, X^{F, IR} \) and \( X^{F, D} \) are four synchronized frames respectively from the top IR, the top depth, the front IR and the front depth cameras. Considering that these images have different modalities, we thus leverage four separate encoders \( \mathcal{E}^{T, IR}, \mathcal{E}^{T, D}, \mathcal{E}^{F, IR} \) and \( \mathcal{E}^{F, D} \) to correspondingly extract spatial features \( f^{T, IR}, f^{T, D}, f^{F, IR} \) and \( f^{F, D} \). To combine them together, there are two distinct avenues: feature-level fusion vs. decision-level fusion.

Figure 2: The architecture of our proposed unimodal DSM model. Pre-trained 2D CNNs are employed as the base encoders to extract features from the input image. Then the channel-wise attention is introduced into these features by leveraging the multi-scale channel attention module (MS-CAM) [30]. Finally, these features are passed to the fully connected (FC) layers to make the prediction \( y_{\text{pred}} \).

Figure 3: The architecture of our proposed multimodal fusion mechanism. Based on the original multi-scale channel attention module (MS-CAM), attentional feature fusion (AFF) and iterative attentional feature fusion (iAFF) [30], we extend the number of attention head in MS-CAM to generate of four weight matrices, as illustrated by (a). Correspondingly, we replace the original MS-CAM in AFF and iAFF with our modified version to allow one-time fusion of four feature maps, as shown by (b) and (c).
Feature-level fusion. The state-of-the-art feature fusion method AFF and iAFF proposed in [30] can only combine two feature maps simultaneously, because their MS-CAM only contains one head, which can at most generate two weight matrices. For example, in Fig. 2, after acquiring one weight matrix \( W \), the other is determined by \( 1 - W \). Hence, to enable the one-time fusion of features from four sources, we extend the number of heads in MS-CAM and use a softmax function to distribute weights, denoted by \( W^T, IR \), \( W^T, D \), \( W^F, IR \) and \( W^F, D \). The structure of our multi-head MS-CAM is illustrated in Figure 3. Then the features fused by AFF are computed by

\[
f' = \sum_{i \in \{T, F\}, j \in \{IR, D\}} f_i, j \otimes W_i, j.
\] (3)

Finally, we feed \( f' \) to a shared FCN \( P \) to make the prediction \( y \).

Decision-level fusion. Similar to the single-modality case, we feed these features to four separate original MS-CAM [30] to introduce global and local attention into each feature. Notice that the attention here is not related to views and modalities. Subsequently, these features are passed to four independent FCNs to calculate scores \( y^T, IR \), \( y^T, D \), \( y^F, IR \) and \( y^F, D \). In the last step, these scores are averaged to collectively make the final prediction:

\[
y = \frac{1}{4} \sum_{i \in \{T, F\}, j \in \{IR, D\}} y_i, j.
\] (4)

3.3 Testing Inferences

Since there are unseen classes in the test set, following [1], we utilize a threshold \( \gamma \) to determine whether the prediction should be "unseen" (\( c_8 \)) or not. The prediction of the class label is determined by

\[
y = \begin{cases} 
c_{\text{arg max } y}, & \text{if } \max y \geq \gamma \\
c_8, & \text{otherwise}
\end{cases}
\] (5)

where \( \max y \) denotes the maximum element of \( y \).

4 Experiments

In this section, we first quantitatively prove the strong similarities between consecutive frames, which questions the essentiality of considering the temporal relations (Sec. 4.1). Then, we demonstrate the effectiveness of our proposed approaches by comparing them with state-of-the-art methods [1] (Sec. 4.2). The results of multiclassification are shown and thoroughly analyzed in Sec. 4.3. As for Sec. 4.4 it illustrates the our models’ real-time inference speeds compared with those in [1], and in Sec. 4.5 the two fusion methods proposed in Sec. 3.2 are compared.

To make fair comparisons with the latest methods, we adopt the 2D versions of the base encoders used in [1], namely, 2D ResNet-18 [16] and 2D MobileNet-V2 [17]. We also feed our models with video clips of 16 frames, with the same spatial data augmentation techniques applied in [1]. We leverage the supervised contrastive loss [31] to fine-tune base encoders \( E \) and the cross-entropy loss to train the prediction layers \( P \). Each model is trained for 50 epochs with a batch size of 128. We employ the Adam algorithm [32] with an initial learning rate of \( 1 \times 10^{-4} \) and a weight decay rate of \( 5 \times 10^{-3} \) as the optimizer. A cosine annealing schedule with warm restarts [33] is exploited to tune the learning rate. Models are
Activity | Histogram Similarity ↑ | Pixel-wise RMSE ↓ | Structural Similarity ↑
|----------------|----------------------|----------------------|----------------------|
| Top IR | Top Depth | Front IR | Front Depth | Top IR | Top Depth | Front IR | Front Depth | Top IR | Top Depth | Front IR | Front Depth |
| Normal driving | 0.99 | 1.00 | 0.97 | 1.00 | 2.75 | 2.18 | 3.40 | 2.02 | 0.96 | 0.84 | 0.95 | 0.88 |
| Talking on the phone | 0.98 | 1.00 | 0.98 | 1.00 | 2.01 | 1.02 | 2.98 | 1.82 | 0.97 | 0.87 | 0.96 | 0.89 |
| Messaging | 0.98 | 1.00 | 0.98 | 1.00 | 2.21 | 2.01 | 2.82 | 1.78 | 0.97 | 0.86 | 0.96 | 0.90 |
| Talking with passengers | 0.98 | 1.00 | 0.98 | 1.00 | 2.06 | 1.97 | 3.14 | 1.87 | 0.98 | 0.86 | 0.96 | 0.89 |
| Reaching behind | 0.98 | 1.00 | 0.97 | 1.00 | 2.05 | 2.02 | 3.46 | 1.90 | 0.97 | 0.86 | 0.95 | 0.88 |
| Adjusting radio | 0.99 | 1.00 | 0.96 | 1.00 | 2.11 | 2.05 | 2.78 | 1.80 | 0.97 | 0.86 | 0.96 | 0.89 |
| Drinking | 0.99 | 1.00 | 0.98 | 1.00 | 2.19 | 2.06 | 2.91 | 1.95 | 0.98 | 0.86 | 0.92 | 0.88 |

Table 4: Average similarities or (differences) under three distinct metrics. The arrows imply whether larger values indicate larger similarities (↑) or not (↓). The histogram and structural similarity should lie between 0 and 1, and being closer to 1 means high similarity. The pixel-wise RMSE is calculated based on the original pixel values in [0, 255]. The high resemblance of neighboring frames of all activities and modalities demonstrates that processing all frames within a video clip is unnecessary.

| Model | AUC-ROC (%) |
|-------|-------------|
|       | Top | Front | Top + Front |
|       | Depth | IR | Depth + IR | Depth | IR | Depth + IR | Depth | IR | Depth + IR |
| MobileNet-V2 [1] | 91.2 | 85.3 | 91.5 | 89.0 | 83.6 | 89.8 | 96.4 | 91.5 | 96.1 |
| ResNet-18 [1] | 91.3 | 88.0 | 91.7 | 90.0 | 87.0 | 92.0 | 96.1 | 93.2 | 96.6 |
| ours (MobileNet-V2 based) | 94.7 | 94.9 | 94.8 | 92.0 | 86.4 | 90.0 | 95.4 | 95.6 | 93.6 |
| ours (ResNet-18 based) | 94.2 | 94.8 | 94.2 | 88.2 | 88.5 | 90.4 | 93.4 | 94.5 | 95.2 |

Table 5: Comparison of models under AUC-ROC in binary classification. In each column (modality), models with the best scores in are indicated in bold typeface, and for each model, the scores of the best modalities are underlined.

implemented in PyTorch and trained on a server with the Ubuntu 20.04 LTS OS and one NVIDIA RTX 3090.

4.1 Frame Similarities

We adopt three common methods to measure how similar two neighboring frames are: the histogram similarity, the mean squared of pixel-wise differences, and the structural similarity [34]. The histogram similarity is based on the difference in the distribution of pixel values. The RMSE is a more rigid metric and can directly show the average pixel differences. The structural similarity is distinct from the former two metrics and is more global and similar to human perceptions. The results based on each activity and modality are shown in Table 4 from which we can see that consecutive frames are highly alike regardless of the choice of metrics. This observation compromises the necessity of considering the temporal dimension, and thus, exploiting 3D convolution on DAD [1] is nonessential.

4.2 Binary Classification

The work [1] is based on binary classification (distinguishing between $c_0$ and $\{c_1, \cdots, c_8\}$). We first compare our models with theirs under the area under the receiver-operating characteristic curve (AUC-ROC) and the accuracy, which are utilized in [1]. Since the test data is imbalanced (66.2% “normal driving” and 33.8% NDRAs), we also leverage the area under the precision-recall curve (AUC-PR) as a more unbiased metric. We adopt feature-level fusion to handle multimodality. Results in terms of AUC-ROC, accuracy, and AUC-PR are shown in Table 5 Table 6 and Table 7 respectively. They confirm that our models can achieve state-of-the-art performance on the DAD [1] dataset (AUC-ROC: 95.6 % vs. 96.6%; accuracy: 92.4% vs. 92.3%).

Table 5 and Table 6 demonstrate that our image-based models can outperform their video-based counterparts in most cases, when using the top view and particularly with the IR channel. In the case of the front view, 3D CNNs are slightly superior to ours only when IR and depth are fused. For the case of combining the top and front views, our proposed methods are again predominant with the IR
### Table 6: Comparison of models in terms of binary classification accuracy. The highest score of each column is indicated in bold and that for each row is also underlined.

| Model                      | Top Accuracy (%) | Front Accuracy (%) |
|----------------------------|------------------|--------------------|
|                            | Depth | IR | Depth + IR | Depth | IR | Depth + IR | Depth | IR | Depth + IR |
| ResNet-18 [1]              | 89.1  | 83.6 | 87.8       | 87.2  | 83.7 | 88.7       | 91.6  | 87.1 | 92.3       |
| Ours (MobileNet-V2 based)  | 91.4  | 92.4 | 91.7       | 87.6  | 81.5 | 85.0       | 90.4  | **90.0** | 89.2       |
| Ours (ResNet-18 based)     | 91.2  | 91.9 | **91.9**   | 85.0  | **84.4** | 86.7       | 88.9  | 88.4 | 90.0       |

Table 7: Comparison of our image-based models under the AUC-PR metric in binary classification. The scores of the best modalities are underlined, and for each given modality, the score of the best model is in shown in bold.

| Modality          | AUC-PR (%) |
|-------------------|------------|
|                   | MobileNet-V2 Based | ResNet-18 Based |
| Top Depth         | 95.7       | 95.8       |
| Top IR            | 95.9       | **96.0**   |
| Top (Depth + IR)  | **96.4**   | 95.7       |
| Front Depth       | 94.3       | 91.9       |
| Front IR          | 91.6       | 91.9       |
| Front (Depth + IR)| 93.9       | 93.0       |
| (Top + Front) Depth| 97.1       | 95.9       |
| (Top + Front) IR  | 97.5       | 96.7       |
| (Top + Front) (Depth + IR)| 96.2| **97.1** |

4.3 Multiclass Classification

Since the classes of DAD [1] are unbalancedly distributed, we sample it during each epoch to ensure that each class has a similar number of instances. We leverage the method illustrated in Sec. 3.1 to predict the unseen class. Results in Table 8 show that for our MobileNet-V2 [17] based approach, similar to the binary class, it still achieves the highest multiclassification accuracy from the top view and the IR modality. However, the DMS based on ResNet-18 [16] benefits more when combining the front and the top views. We visualize the confusion matrices of our ResNet-18-based model [16] in Fig. 5.

The following are important observations:

- **Normal driving.** Our models can adequately recognize this activity in most cases. The highest recall value (98%) is reached when fusing all four modalities.
- **Adjusting radio.** The model based on Top & Depth fails to recognize this action and confuses it with “talking with phone” and “unseen” NDRAs. By comparison, the model based on Top & IR performs much better (79%). The best score (87%) is reached by the model based on (Top + Front) with IR. For the front view, fusing modalities can lead to better results, and similarly, for the IR modality, the combination of views provides higher scores.
- **Drinking.** The model based on Top & Depth also fails to recognize the “drinking” activity with sufficient accuracy and misclassifies a significant pro-
Figure 5: The confusion matrices of our ResNet-18-based model on DAD. There are $3 \times 3$ plots in this figure. Each row corresponds to a stream, and each column represents a view. The row of each confusion matrix is the ground-truth, while the column of each matrix is the prediction. Matrices are normalized to have rows summing up to 1, hence diagonal entries represent the recall values.

| Depth | Normal Driving | Adjusting Radio | Messaging | Reaching Behind | Talking with Passenger | Unseen NDRAs |
|-------|----------------|-----------------|-----------|-----------------|-----------------------|--------------|
| Top   | 0.94           | 0.00            | 0.00      | 0.00            | 0.05                  |              |
| Front | 0.00           | 0.00            | 0.00      | 0.00            | 0.00                  |              |
| Top + Front | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |              |

| Depth + IR | Normal Driving | Adjusting Radio | Messaging | Reaching Behind | Talking with Passenger | Unseen NDRAs |
|------------|----------------|-----------------|-----------|-----------------|-----------------------|--------------|
| Top   | 0.92           | 0.07            | 0.07      | 0.07            | 0.07                  |              |
| Front | 0.07           | 0.07            | 0.07      | 0.07            | 0.07                  |              |
| Top + Front | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 |              |

| Depth + IR | Normal Driving | Adjusting Radio | Messaging | Reaching Behind | Talking with Passenger | Unseen NDRAs |
|------------|----------------|-----------------|-----------|-----------------|-----------------------|--------------|
| Top   | 0.95           | 0.05            | 0.05      | 0.05            | 0.05                  |              |
| Front | 0.05           | 0.05            | 0.05      | 0.05            | 0.05                  |              |
| Top + Front | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |              |

| Depth + IR | Normal Driving | Adjusting Radio | Messaging | Reaching Behind | Talking with Passenger | Unseen NDRAs |
|------------|----------------|-----------------|-----------|-----------------|-----------------------|--------------|
| Top   | 0.95           | 0.05            | 0.05      | 0.05            | 0.05                  |              |
| Front | 0.05           | 0.05            | 0.05      | 0.05            | 0.05                  |              |
| Top + Front | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |              |

| Depth + IR | Normal Driving | Adjusting Radio | Messaging | Reaching Behind | Talking with Passenger | Unseen NDRAs |
|------------|----------------|-----------------|-----------|-----------------|-----------------------|--------------|
| Top   | 0.95           | 0.05            | 0.05      | 0.05            | 0.05                  |              |
| Front | 0.05           | 0.05            | 0.05      | 0.05            | 0.05                  |              |
| Top + Front | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |              |

| Depth + IR | Normal Driving | Adjusting Radio | Messaging | Reaching Behind | Talking with Passenger | Unseen NDRAs |
|------------|----------------|-----------------|-----------|-----------------|-----------------------|--------------|
| Top   | 0.95           | 0.05            | 0.05      | 0.05            | 0.05                  |              |
| Front | 0.05           | 0.05            | 0.05      | 0.05            | 0.05                  |              |
| Top + Front | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |              |
portion (45%) as “unseen”. By comparison, when also incorporating the IR modality, the results can be much better improved, with the highest recall score of 53%. For other cases, combining views or modalities may not enhance the performance.

- **Messaging.** This is an interesting case, as we find that except for Top & IR, models based on other modalities are confused between “messaging” and “normal driving”. This confusion represents a critical noise factor that could potentially cause severe consequences. Another finding is that the top view is more informative in recognizing this action.

- **Reaching behind.** Again, the model based on Top & Depth cannot correctly classify this action. This action can be better observed from the front view, which is intuitive. The largest recall value (76%) is obtained when using (Top + Front) & Depth.

- **Talking with passenger.** Similar to “reaching behind” and as shown in Fig. 1, this activity is only recognizable from the front view, but the front view model, unfortunately, achieves a recall score of only (42%). We believe the reason for this substandard performance is that the camera resolution is not high enough to capture the lip movements.

- **Talking with phone.** Some models confuse this activity with “normal driving” and “unseen” NDRAs. Although the model based on Top & IR does not have the highest recall value, it appears to be capable to distinguish between this action and “normal driving” reasonably well.

- **Unseen.** Models based on Top & IR and Top & Depth have more nontrivial values (56% and 37%, respectively), while others seem incapable to recognize this class well. To summarize, although our models perform very well on the binary classification task, they cannot achieve similar performance on multiclassification. Among all activities, our models can recognize “normal driving” well but can be confused with other classes. We attribute the low recall values of NDRAs to the limited representation power of ResNet-18 and the extreme imbalanced class distribution in the training set of DAD. On the other hand, we think overall the top view and the IR modality is more informative, while by contrast, the models based on Top & Depth have suboptimal performance.

4.4 **Computational Efficiency**

We compare our models with those leveraged in [1] in terms of the number of floating-point operations (FLOPs) and the inference time. Each model is tested 1,000 times to reduce the effect of stochasticity. During each test-run, a randomly generated matrix of size $(16, 171, 224)$, which simulates a normalized video clip, is fed to each model. We select two platforms to represent different scenarios:

- An M1 Pro chip (ARM64) with an 8-core CPU ($6 \times 3.2$ GHz) and a 14-core GPU (1.30 GHz), and a 16 GB RAM to simulate embedded devices in cars.
- An AMD 5900x (AMD64) with a 12-core CPU (3.7 GHz) and a 64 GB RAM to simulate cloud servers.

Code is implemented in Python 3.9.13 and PyTorch 1.12.1. Note that the Infineon CamBoard pico flexx cameras used during data collection in DAD have a frame rate of 45 Hz. A model should then complete the inference before the next video clip is collected and fed to it. Therefore, real-time models should take no longer than $16/45 \approx 0.3556$ seconds to infer each video clip. Table 9 presents a detailed comparison of models. From it we can see that all our image-based models can make pre-
dictions within the expected time and are 10 times faster than those that use 3D CNNs, demonstrating exceptional efficiencies.

4.5 Comparison of Fusion Strategies

To handle multiple modalities, we utilize mainly two fusion mechanisms: feature-level fusion and decision-level fusion, as elaborated in Sec. 3.2. Table 10 compares them with the base encoder ResNet-18 [16], from which we can see that feature-level fusion outperforms decision level fusion when fusing two views. As for other cases, performance gaps are not significant.

| Modality                  | AUC-PR (%) |
|---------------------------|------------|
|                           | Feature Level | Decision Level |
| Top (Depth + IR)          | 95.7        | 96.2            |
| Front (Depth + IR)        | 93.0        | 93.4            |
| (Top + Front) Depth       | 95.9        | 95.6            |
| (Top + Front) IR          | 96.7        | 95.1            |
| (Top + Front) (Depth + IR)| 97.1        | 95.7            |

Table 10: The AUC-PR values of the two fusion methods with base encoder ResNet-18 [16]. For each modality, the higher score is in bold.

5 Conclusion

Based on our experiments, we discovered that compared with video-based models, our image-based methods can achieve similar performances and are exceedingly more efficient. Therefore, we can conclude that employing image-based models for the real-time inference of driver activities is more practical and less computationally complex. As part of this work, we also augmented the labelling of the DAD test set for future use by the scientific community. We found that data from the top view and the infrared stream are more informative and enhance model performance as confirmed by comparatively higher scores for binary classification and negligible confusion between phone-involved activities and normal driving. These results demonstrate a good practical value. Our models, however, do not perform the same on the multiclassification of NDRTs as on the binary classification problem. We ascribe this issue to the limited representation power of the backbones and to data imbalance. In the future, we plan to replace the backbone with the state-of-the-art classifiers and leverage more advanced techniques to improve multiclassification performance.

References

[1] O. Köpüklü, J. Zheng, H. Xu, and G. Rigoll, “Driver anomaly detection: A dataset and contrastive learning approach,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 91–100.

[2] K. Diaz-Chito, A. Hernández-Sabaté, and A. M. López, “A reduced feature set for driver head pose estimation,” Applied Soft Computing, vol. 45, pp. 98–107, 2016.

[3] Q. Massoz, T. Langohr, C. François, and J. G. Verly, “The ulg multimodality drowsiness database (called drozy) and examples of use,” in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2016, pp. 1–7.

[4] M. Roth and D. M. Gavrila, “Dd-pose-a large-scale driver head pose benchmark,” in 2019 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2019, pp. 927–934.

[5] A. Schwarz, M. Haurilet, M. Martinez, and R. Stiefelhagen, “Driveahead—a large-scale driver head pose dataset,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1–10.

[6] C.-H. Weng, Y.-H. Lai, and S.-H. Lai, “Driver drowsiness detection via a hierarchical temporal deep belief network,” in Asian Conference on Computer Vision, Springer, 2016, pp. 117–133.

[7] N. Das, E. Ohn-Bar, and M. M. Trivedi, “On performance evaluation of driver hand detection algorithms: Challenges, dataset, and metrics,” in 2015 IEEE 18th international conference on intelligent transportation systems, IEEE, 2015, pp. 2953–2958.
[8] O. Köpükli, T. Ledwon, Y. Rong, N. Kose, and G. Rigoll, “Drivermhg: A multi-modal dataset for dynamic recognition of driver micro hand gestures and a real-time recognition framework,” in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, 2020, pp. 77–84.

[9] E. Ohn-Bar, S. Martin, and M. Trivedi, “Driver hand activity analysis in naturalistic driving studies: Challenges, algorithms, and experimental studies,” Journal of Electronic Imaging, vol. 22, no. 4, p. 041119, 2013.

[10] Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa, “Real-time distracted driver posture classification,” in Neural Information Processing Systems (NIPS 2018), Workshop on Machine Learning for Intelligent Transportation Systems, 2018.

[11] M. Martin, A. Roitberg, M. Haureit, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen, “Drive&Act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2801–2810.

[12] J. D. Ortega, N. Kose, P. Cañas, M.-A. Chao, A. Unnervik, M. Nieto, O. Otaegui, and L. Salgado, “Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis,” in European Conference on Computer Vision, Springer, 2020, pp. 387–405.

[13] H. M. Eraqi, Y. Abouelnaga, M. H. Saad, and M. N. Moustafa, “Driver distraction identification with an ensemble of convolutional neural networks,” Journal of Advanced Transportation, vol. 2019, 2019.

[14] B. Baheti, S. Gajre, and S. Talbar, “Detection of distracted driver using convolutional neural network,” in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 1032–1038.

[15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in International Conference on Learning Representations, 2015.

[16] 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, 2016, pp. 770–778.

[17] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[18] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[19] O. Köpükli, N. Kose, A. Gunduz, and G. Rigoll, “Resource efficient 3d convolutional neural networks,” in 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, 2019, pp. 1910–1919.

[20] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “Shufflenet v2: Practical guidelines for efficient cnn architecture design,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 116–131.

[21] M. Gutmann and A. Hyvärinen, “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models,” in Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[22] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[25] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in Thirty-first AAAI conference on artificial intelligence, 2017.

[26] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.

[27] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[28] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[29] X. Li, W. Wang, X. Hu, and J. Yang, “Selective kernel networks,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 510–519.

[30] Y. Dai, F. Gieseke, S. Oehmecke, Y. Wu, and K. Barnard, “Attentional feature fusion,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3560–3569.

[31] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Supervised contrastive learning,” Advances in Neural Information Processing Systems, vol. 33, pp. 18 661–18 673, 2020.

[32] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[33] I. Loshchilov and F. Hutter, “Sgdr: Stochastic gradient descent with warm restarts,” arXiv preprint arXiv:1608.03983, 2016.