Weighted Multi-Task Vision Transformer for Distraction and Emotion Detection in Driving Safety

Yixiao Wang, Zhe Li, Guowei Guan, Yipin Sun, Chunyu Wang, Hamid Reza Tohidypour,
Panos Nasiopoulos, and Victor C.M. Leung
Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, Canada
yixiaow@ece.ubc.ca, {lizhe918, gguan3, yipinsun, shycw1}@student.ubc.ca, {htohidyp, panosn, vleung}@ece.ubc.ca

***Abstract*—Detecting drivers' distraction and emotion has raised attention due to its importance in ensuring driving safety, especially with the increasing number of accidents caused by distracted or emotionally unstable drivers. Previous research has employed the multi-tasking method to detect these two factors simultaneously but paid insufficient attention to the emotion detection part. Meanwhile, existing publicly available datasets use side cameras to capture both distraction and emotion, which is impractical in real driving scenarios. To address these issues, in this paper we propose a vision transformer-based approach that enhances the emotion detection accuracy while maintaining the performance of distraction detection by balancing the two. Moreover, we generated a new dataset that includes a front view of the driver's face, which improves the accuracy of emotion detection. Evaluation results validate the effectiveness of the proposed approach and demonstrate the balanced importance of emotion detection in driving safety. Our proposed method presents valuable contributions towards enhancing the safety of driving by highlighting the significance of emotion detection and introducing practical solutions to improve its accuracy.**

***Index Terms*—emotion detection, distraction detection, multi-tasking, vision transformer, penalty weight**

I. Introduction

The detection of driver distraction and emotion has emerged as a critical factor in maintaining road safety. It is widely known that the behavior of a driver significantly affects the accident rates on the road. Distraction is a primary cause of accidents and can result from several factors, such as the use of mobile phones, eating, or conversing with passengers. Fig. 1 has shown several typical distraction and emotion examples. This behavior has been identified as a leading cause of accidents that result in fatalities or severe injuries. Thus, identifying such distractions and alarming the driver can help to reduce the number of car accidents. Apart from distraction, emotions can also have a significant impact on driver behavior and decision-making. Strong emotional responses such as anger, anxiety, or fatigue can impair the driver's ability to react appropriately to various road situations [1]. For example, an angry or anxious driver is likely to be more aggressive on the road, increasing the risk of accidents. Fatigue is another emotion that can lead to driver distraction, as it affects the driver's focus and alertness. Therefore, identifying and addressing both driver distraction and emotion is crucial in ensuring road safety. The use of advanced technologies, such as sensors and cameras, can aid in detecting and analyzing driver behavior, alerting drivers to potential distractions or emotional distress.

(a) Typical distraction behavior of drivers, including texting, talking on the phone, operating the radio, drinking, reaching behind, talking to the passengers [5].

(b) Typical emotion during driving, including neural, angry, happy, sad, and surprise [13].

Fig. 1: Examples of typical distraction and emotion of the drivers.

In recent years, there has been a surge of research on detecting distractions or emotions in drivers. For instance, Zhang et al., proposed a connected CNN that fuses low-level and high-level features to recognize facial expressions, achieving 97% accuracy on the KMU FED dataset [2]. Another approach proposed by Leone et al., was based on VGG16 and uses facial expression recognition to detect driver road rage, achieving 94.27% accuracy on the KMU-FED dataset [3]. More recently, Ma et al. have introduced a novel model, ViT-DD, that utilizes the vision transformer architecture for simultaneous detection of driver distraction and emotion [4]. This model is designed to work with body and face images of the driver. The researchers

979-8-3503-7099-7/24/$31.00 ©2024 IEEE

trained and evaluated ViT-DD using the State Farm Distracted Driver Detection (SFDDD) dataset [5] and the American University in Cairo Distracted Driver Dataset (AUCDD) [6], which both contain side-view driver frames. The authors were able to achieve promising results in detecting both driver distraction and emotion simultaneously. Although ViT-DD is presently considered state-of-the-art and offers the capability to detect both driver distraction and emotion simultaneously, the model's performance in detecting emotions is not entirely satisfactory. This issue can be attributed to two primary factors. Firstly, the side-view images used as input to the network do not provide sufficient facial information for accurate emotion prediction. Secondly, compared to distraction, the facial area that represents emotion occupies a relatively small fraction of the input frame. This makes it challenging for the model to distinguish and interpret subtle facial expressions accurately.

In this paper, we introduce a new approach to address the above identified challenges in detecting driver distraction and emotion. Our approach employs a novel weighting scheme that balances the importance of distraction and emotion detection. Specifically, we assign higher weight to emotion detection during the training process, which encourages the model to give more emphasis to detecting emotions in addition to distraction. By using this approach, we achieve improved accuracy in detecting driver emotions without compromising the performance of distraction detection. Furthermore, we constructed a dataset for training and evaluation of our approach. By employing front-view captured face images, which is more aligned with real-life scenarios, our proposed dataset provides additional information about the driver's facial expressions, enabling the model to learn and interpret subtle emotional cues more accurately. Performance evaluations have shown that our method improves the emotion detection accuracy by 9% compared to state-of-the-art. The rest of the paper is structured as follows. Section 2 gives an overview of existing related work. In Section 3, we describe our proposed approach and in Section 4 we present the evaluation process and discuss the results. Section 5 concludes the paper.

II. Related Work

A. Facial emotion detection

Facial emotion detection has seen significant development in recent years, thanks to advancements in computer vision and machine learning. Regular facial emotion detection systems can accurately recognize a wide range of emotions based on facial expressions, including happiness, sadness, anger, fear, and surprise. These systems have a wide range of applications, such as social media platforms, e-commerce, and customer service. Several facial recognition frameworks have been developed to enable emotion detection [7-8]. However, detecting emotions in driving scenarios presents unique challenges that require specialized solutions. In general, facial emotion detection models are trained on datasets containing a wide range of facial expressions captured in various contexts, such as in the lab, at home, or in public places. In contrast, driving scenarios are characterized by specific and often repetitive events, such as lane changes, intersections, or traffic congestion, that can trigger specific emotions in drivers. Therefore, facial emotion detection models for driving scenarios need to be trained on specialized datasets that reflect these unique driving situations.

Furthermore, common facial emotion detection models may not be able to detect subtle emotional cues that are relevant to driving scenarios, such as frustration, fatigue, or anxiety, which can negatively impact driver behavior and decision-making. Thus, specialized facial emotion detection models for driving scenarios need to be designed to capture these subtle emotional cues that can affect driving safety.

B. Multi-tasking of distraction and emotion detection

Multi-tasking models for distraction and emotion detection, such as ViT-DD, have become increasingly important for driver safety. These models use advanced computer vision techniques and machine learning algorithms to analyze both facial expressions and body language of drivers, simultaneously detecting signs of distraction and emotion. ViT-DD is based on the Vision Transformer model, and can detect both distraction and emotion by analyzing body and face images of the driver, using datasets such as the SFDDD and AUCDD. This network takes a pair of images as input, the first showing the side view of driver's body and the second showing side view of their face. As a multi-task model, ViT-DD generates two simultaneous predictions for each image pair: one for the distraction level of the driver and another for their emotion. By simultaneously detecting both distraction and emotion, these multi-tasking models offer significant potential to improve road safety and reduce the number of accidents caused by emotional driving.

However, as mentioned above, ViT-DD is trained on two datasets: SFDDD and AUCDD, both of which only captured the side-view images of drivers. In addition, ViT-DD's overly heavy emphasis on distraction leads to unsatisfactory performance on emotion detection.

III. Our Proposed Method

A. Vision transformer based multi-tasking network

In this paper, we propose a vision transformer-based network to detect distraction and emotion together. The network in our proposed method is shown in Fig. 2. We choose the ViT-DD network as the baseline, which is known as the state-of-the-art multi-tasking network for driver distraction and emotion detection. Vision transformer (ViT) is a novel architecture for image recognition that uses a transformer-like model to process patches of an image [9]. Unlike convolutional neural networks (CNNs), which rely on local features and spatial correlations, ViT learns global dependencies and semantic relationships among image patches using self-attention mechanisms. This network has shown remarkable performance and efficiency on various vision tasks, such as image classification, segmentation, detection, and retrieval. It can also benefit from large-scale pre-training on unlabeled images, which can improve its generalization and robustness.

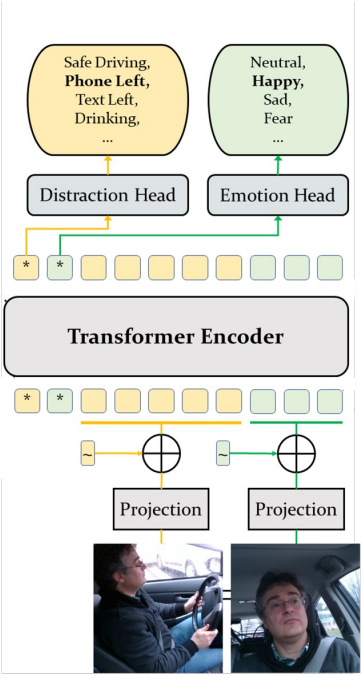

Fig. 2: Example of a figure caption.

However, the difficulties in recognizing and analyzing facial expressions are still severe due to the low resolution of the emotional features. Unlike other tasks that require attention to the whole input frame, such as object detection or scene classification, facial expression recognition focuses on a small region of the face that conveys the emotional state. This region often occupies only a fraction of the input frame, and may be affected by factors such as illumination, occlusion, pose, and distance. Therefore, it is challenging for the model to capture and interpret the subtle variations in facial expressions that reflect different emotions.

To overcome this challenge, we introduce a two-input network that combines the side view of the body and the front view of the face as complementary sources of information. The side view of the body provides global context and posture information, while the front view of the face provides fine-grained details and facial expressions. The two inputs are first embedded and concatenated, and then sent to the vision transformer. The vision transformer splits the input image into patches of fixed size (e.g., 16x16 pixels), and treats each patch as a token. It then applies multiple layers of self-attention and feed-forward networks to encode the relationships between patches. The output of the vision transformer is a sequence of feature vectors, one for each patch. The network has two detection heads, one for distraction and the other for emotion. The distraction head aims to classify whether the person is distracted or not, based on the global features of the body and the face. The emotion head aims to classify the person's emotion into one of six categories (happy, sad, angry, surprised, disgusted, or neutral), based on the local features of the face. Both heads use a linear layer followed by a softmax layer to produce the final predictions.

B. Dataset generation

To demonstrate the effectiveness of our approach, we create a large-scale dataset that contains front and side view images with corresponding annotations. We use this dataset to train and evaluate our method for detecting distraction and emotion of drivers. Our dataset is based on the Driver Monitoring Dataset (DMD), which contains videos of drivers performing various tasks and their corresponding distraction labels [10]. The DMD dataset is very large, so we only select a subset of it for our experiments. More specifically, we chose the videos of four drivers: gA1, gA2, gB6, and gB7. These drivers have different genders and ages, which can help us evaluate the generalization ability of our model. For each driver, we use the first four video sets, which are labeled with distraction events such as phone calls, texting, eating, drinking, etc. The other video sets are not relevant for our task, so we ignore them. In each video set, we have two types of videos: RGB front-view and RGB side-view. These videos are recorded simultaneously from different angles, which can provide us with rich information about the driver's head pose, facial expression, eye gaze, etc. We use these two videos as a pair of inputs for our model.

In order to process the video pairs from DMD, we perform several preprocessing steps. Firstly, we reduce the number of redundant frames in both the side and front-view videos by selecting every tenth frame and aligning them based on their frame number. Secondly, we utilize RetinaFace [11], an advanced face detection algorithm, to identify and extract the face regions from both the side and front-view images. Finally, we use PAZ [12], a deep learning framework specifically designed for facial expression recognition, to assign emotion labels to each cropped front-view face image. Our emotion classification system utilizes the six basic emotions, namely anger, disgust, fear, happiness, sadness, and surprise, as the emotion categories. By employing these preprocessing steps, we are able to extract relevant and accurate data from the DMD videos for our subsequent analyses. The use of RetinaFace and PAZ allows for efficient and precise detection and labeling of facial expressions, ensuring the validity of our results. At last, we adopt the ground-truth distraction labels provided by DMD, resulting in a total of 9505 frames. Then, the processed frames are split into training and validation sets in a ratio of 9:1.

C. Weight coefficient design

Our network consists of two branches: a distraction detection branch and an emotion detection branch. The two branches are based on the ViT architecture, which can capture global and local features from images. The two branches share the same ViT encoder, which enables knowledge transfer and feature fusion between the two tasks. This encoder takes two inputs: the driver's body image and face image. The body image is used to detect the driver's posture and head orientation, which are indicators of distraction. The face image is used to detect the driver's facial expression, which is an indicator of emotion. The two branches output the distraction

and emotion labels respectively. This approach leads to the choice of a compound loss function L, and here we choose the similar weighted-sum structure, as follows:

$$L = \lambda_1 * L_{distration} + \lambda_2 * L_{emotion} \quad (1)$$

where the $L_{distration}$ and the $L_{emotion}$ are the loss value from the distraction and emotion branch, respectively, and λ_1 and λ_2 are the corresponding coefficients. Here, the two weight coefficients λ_1 and λ_2 sum to 1. That is, increasing λ_1 decreases λ_2 and vice versa.

In this context, the emotion weight coefficient is a parameter that controls the relative importance of the emotion detection task compared to the distraction detection task. The higher the emotion weight coefficient, the more attention the model pays to the emotion detection task. When a higher weight coefficient λ_1 is assigned to distraction, the network prioritizes the correct detection of distraction, as an incorrect prediction results in a heavier penalty and a bigger loss, which the network aims to minimize. Similarly, assigning a higher weight coefficient to emotion motivates the network to prioritize accurate emotion predictions, as higher losses and penalties are applied for incorrect emotion detection. Conversely, a smaller weight coefficient on distraction or emotion reduces the priority of the network to optimize for that task.

In the original ViT-DD approach, the primary goal is distraction detection, which influences the design of its loss function. Specifically, the weight coefficient for distraction λ_1 is set to a high value of 0.999, with the weight coefficient for emotion λ_2 set to 0.001. This configuration indicates that ViT-DD receives almost no penalty for incorrect emotion prediction during training and, therefore, does not prioritize emotion detection. Consequently, the dissatisfaction performance of ViT-DD on emotion detection can be attributed to this weight coefficient configuration.

In our proposed method, we recognize the significance of determining the optimal weight coefficients for both distraction and emotion in a multi-tasking network. The relationship between these two branches underscores the importance of finding the right balance. Through rigorous experimentation, we identified the weight coefficients that yield the best performance for both tasks. The optimization of these coefficients becomes particularly crucial as it allows us to effectively leverage the interconnected nature of the network. By finding the optimal weight coefficients, we enhance the overall performance and ensure that both distraction and emotion detection receive the appropriate attention and prioritization. Details are presented in the following Section.

IV. Experimental Results

In this paper, we adopt both SFDDD and our generated DMD dataset for training and evaluation. SFDDD is a large-scale dataset for driver distraction detection, containing 22,424 images of drivers performing 10 different activities, such as texting, talking on the phone, or adjusting the radio. It was captured by a constant-placed 2D dashboard camera with 640 × 480 pixels in RGB. For the SFDDD dataset, we use the original side view frames and the emotion labels from the original ViT-DD. The SFDDD dataset has two split ways, one is split-by-image (sbi) where all the images are randomly assigned to training and validation set by a certain ratio, while the other method is split-by-driver (sbd) where all the images of certain drivers are assigned to the training set while the images of other drivers are assigned to the validation set. For DMD dataset, after applying the preprocessing procedure on the original DMD dataset described in Section 3.2., it results in two sets of configurations: DMD (Side+Front) and DMD (Side+Side). The DMD (Side+Front) consists of images of the body from the side view and images of the face from the front view, while the DMD (Side+Side) contains images of both the body and the face from the side view. These two sets have the same labels for distraction and pseudo-emotion for fair comparison in our further analysis.

Initially, we conducted experiments on the SFDDD dataset to evaluate our method and compare it with the original ViT-DD. Given that the original paper lacked ViT-DD's emotion detection accuracy, we reproduced the results using the original ViT-DD's weight coefficient to establish a fair baseline for comparison. This allowed us to objectively assess our approach against the baseline. Following the same setup as the original ViT-DD paper, including datasets and evaluation metrics, we present our method's experimental results in Table I. The original ViT-DD achieved 0.9963 accuracy for distraction detection on the sbi split of SFDDD and 0.9251 on the sbd split. When we adopted the same weight coefficient configuration (0.999 for distraction and 0.001 for emotion), our method exhibited similar performance. However, noteworthy improvements in emotion detection emerged when we increased the emotion weight coefficient to 0.25. Emotion detection accuracy on the sbi split rose from 0.8014 to 0.8692, while on the sbd split, it improved from 0.7436 to 0.8417. These results underscore our method's ability to balance weights effectively between distraction and emotion detection, leading to enhanced emotion detection in driver monitoring.

We extended our evaluation to the DMD dataset, considering both Side+Side (SS) and Side+Front (SF) configurations, addressing the lack of front-view driver face data in SFDDD. Our network training with various weight coefficients yielded results in Table II. Initially, our method achieved a distraction accuracy of 0.9435 and an emotion accuracy of 0.6406 on the DMD (SS) setup. Incorporating front-view face images improved emotion accuracy to 0.6463 while maintaining distraction accuracy at 0.9368. This demonstrates our approach's effectiveness in enhancing emotion detection without compromising distraction detection. Front-view images provide richer cues for capturing facial expressions, especially around the eyes and mouth, essential for emotion detection. Section 3.3 highlights our network retraining, exploring different coefficients to balance distraction and emotion detection, with four rounds of experimentation using varying coefficients.

Table II presents our approach's evaluation results in the

TABLE I: Performance comparison for different network on SFDDD with different coefficients, the top and second are marked as bold and bold italic, respectively.

Dataset		Coefficients		Accuracy	
		λ_1	λ_2	Distraction	Emotion
SFDDD (sbi)	ViT-DD	0.999	0.001	0.9963	-
	Ours	0.999	0.001	***0.9963***	***0.8014***
	Ours	0.75	0.25	**0.9981**	**0.8692**
SFDDD (sbd)	ViT-DD	0.999	0.001	**0.9251**	-
	Ours	0.999	0.001	***0.9218***	0.7436
	Ours	0.75	0.25	0.9125	**0.8417**

last four rows. Adjusting the coefficients for distraction and emotion detection significantly influenced both tasks. When the emotion coefficient rose to 0.25 and the distraction coefficient fell to 0.75, emotion detection accuracy increased to 0.7326, with distraction accuracy also improving to 0.9411. This improvement stems from the network learning shared relevant features for multiple tasks, enhancing generalization to new data. Setting coefficients at 0.5 for both tasks maintained emotion detection accuracy at 0.7326, while distraction accuracy slightly decreased to 0.94. A further reduction of the distraction coefficient to 0.25 lowered accuracy in both tasks, with emotion accuracy at 0.7126 and distraction at 0.9368. Lastly, a distraction coefficient of 0.001 and an emotion coefficient of 0.999 raised emotion detection accuracy slightly to 0.7253 but drastically reduced distraction accuracy to 0.8705.

In our experiments, we explored how the emotion weight coefficient influences multi-task learning's performance in distraction and emotion detection. Increasing this coefficient enhanced both tasks, suggesting mutual benefits between them. However, excessive emotion emphasis (0.999) and minimal distraction focus (0.001) decreased accuracy in both tasks, indicating an imbalance. The model became overly fixated on emotion detection, neglecting distraction detection due to limited features from small input face images. Therefore, we conclude that there exists an optimal emotion weight coefficient range. Our experiments revealed that a coefficient of 0.25 yielded the best results for both distraction and emotion detection tasks.

V. Conclusion

In conclusion, in this paper we presented a novel approach for detecting distraction and emotion in driving scenarios. The proposed vision transformer-based approach effectively balances the importance of both factors and improves the accuracy of emotion detection by incorporating the front view of the driver's face. The newly generated dataset enables the use of practical front-view cameras for emotion detection, which is not feasible in existing datasets that rely solely on side cameras. The experimental results demonstrate the effectiveness of the proposed method and validate its contribution to driving safety. By emphasizing the importance of emotion detection and introducing a practical solution to enhance its accuracy, this paper contributes to the research on improving the safety of driving. The proposed approach and dataset have practical implications for the development of advanced driver assistance systems that can effectively detect and respond to both distraction and emotion in real-time.

TABLE II: Performance comparison for our proposed network in different view and coefficient configurations, the top and second are marked as bold and bold italic, respectively. SS means Side+Side set of DMD while SF means Side+Front set of DMD.

Dataset		Coefficients		Accuracy	
		λ_1	λ_2	Distraction	Emotion
DMD (SS)	Ours	0.999	0.001	**0.9435**	0.6406
DMD (SF)	Ours	0.999	0.001	0.9368	0.6463
	Ours	0.75	0.25	***0.9411***	**0.7326**
	Ours	0.5	0.5	0.94	***0.7326***
	Ours	0.25	0.75	0.9368	0.7126
	Ours	0.001	0.999	0.8705	0.7253

References

[1] S. B. Sukhavasi, S. B. Sukhavasi, K. Elleithy, A. El-Sayed, and A. Elleithy, "A hybrid model for driver emotion detection using feature fusion approach," International journal of environmental research and public health, vol. 19, no. 5, p. 3085, 2022.

[2] Zhang, J.; Mei, X.; Liu, H.; Yuan, S.; Qian, T. Detecting negative emotional stress based on facial expression in real time. In Proceedings of the 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), Wuxi, China, 19–21 July 2019; IEEE: New York, NY, USA, 2019; pp. 430–434.

[3] Leone, A.; Caroppo, A.; Manni, A.; Siciliano, P. Vision-based road rage detection framework in automotive safety applications. Sensors 2021, 21, 2942.

[4] Y. Ma and Z. Wang, "ViT-DD: Multi-Task Vision Transformer for Semi-Supervised Driver Distraction Detection."

[5] 'State Farm Distracted Driver Detection', Kaggle, 2023. [Online]. Available: https://www.kaggle.com/c/state-farm-distracted-driver-detection. [Accessed: Apr. 26, 2023]

[6] 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.

[7] N. Reddy and R. Derakhshani, 'Emotion Detection using Periocular Region: A Cross-Dataset Study', in 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, United Kingdom: IEEE, Jul. 2020, pp. 1–6.

[8] S. B. Sukhavasi, S. B. Sukhavasi, K. Elleithy, A. El-Sayed, and A. Elleithy, 'A Hybrid Model for Driver Emotion Detection Using Feature Fusion Approach', IJERPH, vol. 19, no. 5, p. 3085, Mar. 2022

[9] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: International Conference on Learning Representations. 2020.

[10] J. D. Ortega et al., "Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, 2020: Springer, pp. 387-405.

[11] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild. arXiv 2019," arXiv preprint arXiv:1905.00641, 1905.

[12] O. Arriaga, M. Valdenegro-Toro, M. Muthuraja, S. Devaramani, and F. Kirchner, "Perception for autonomous systems (paz)," arXiv preprint arXiv:2010.14541, 2020.

[13] H. Xiao et al., 'On-Road Driver Emotion Recognition Using Facial Expression', Applied Sciences, vol. 12, no. 2, p. 807, Jan. 2022.