Leveraging Vision Language Models for Facial Expression Recognition in Driving Environment

Ibtissam Saadi
Faculty 1 MINT, Brandenburg University of Technology BTU Cottbus-Senftenberg
Cottbus, Germany
ibtissam.saadi@b-tu.de

Abdenour Hadid
Sorbonne Center for Artificial Intelligence, Sorbonne University Abu Dhabi
Abu Dhabi, UAE
abdenour.hadid@ieee.org

Douglas W. Cunningham
Faculty 1 MINT, Brandenburg University of Technology BTU Cottbus-Senftenberg
Cottbus, Germany
douglas.cunningham@b-tu.de

Abdelmalik Taleb-Ahmed
Laboratory of IEMN, CNRS, Centrale Lille, UMR 8520, Univ. Polytechnique Hauts-de-France
Valenciennes, France
abdelmalik.taleb-ahmed@uphf.fr

Yassin El Hillali
Laboratory of IEMN, CNRS, Centrale Lille, UMR 8520, Univ. Polytechnique Hauts-de-France
Valenciennes, France
Yassin.ElHillali@uphf.fr

Abstract

We are witnessing an increasing interest in vision-language models (VLMs) as reflected in the impressive results across a large spectrum of tasks. In this context, we introduce in this paper a novel architecture that exploits the capabilities of VLMs for facial expression recognition in driving environment to enhance road safety. We present an approach called CLIVP-FER, which uses the Contrastive Language-Image Pretraining (CLIP) and combines both visual and textual data to overcome the environmental challenges and ambiguities in facial expression interpretation. In addition, we apply average pooling to improve the accuracy and the computational efficiency. The proposed approach is thoroughly evaluated on a benchmark driving dataset called KMU-FED. The experiments showed superior performance compared to state-of-the-art methods, achieving an average accuracy of 97.36%. Cross-database evaluation is also provided showing good generalization abilities. The ablation study gives more insights into the performance of our proposed architecture. The obtained results are interesting and confirm the capabilities of vision-language models in vision tasks, demonstrating their promising applications in efficient driver assistance and intervention systems. We are making the code of this work publicly available for research purposes at https://github.com/Ibtissam-SAADI/CLIVP-FER.

CCS Concepts

• **Computing methodologies** → **Computer Vision**; **Activity recognition and understanding**.

iWOAR 2024, September 26–27,2024, Postdam, Germany

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Keywords

Facial Expression Recognition, Driver's Emotions, Vision Language Models, Contrastive Language-Image Pretraining

ACM Reference Format:
Ibtissam Saadi, Abdenour Hadid, Douglas W. Cunningham, Abdelmalik Taleb-Ahmed, and Yassin El Hillali. 2024. Leveraging Vision Language Models for Facial Expression Recognition in Driving Environment. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Road safety is a major concern for the automotive industry. Human error is a significant contributor to road accidents, resulting in various safety issues. Technological advancements have led to the development of systems aimed at enhancing the driving experience and reducing accident rates. Specifically, facial expression recognition (FER) systems are increasingly being integrated into autonomous vehicles and advanced driver assistance systems (ADAS), as illustrated in Figure 1. These systems are crucial for detecting emotional states in driving environment [18] [14].

In this context, numerous studies have investigated methodologies for recognizing drivers' facial expressions, ranging from using handcrafted features to deep learning such as [15], [8], [24], [6], and [17]. These systems typically depend mainly on visual data to analyze facial cues and infer emotions. However, several factors may significantly affect the accuracy of these systems, including variable lighting conditions and uncooperative users. Moreover, an over-reliance on visual data can restrict the contextual interpretation of emotions, possibly leading to errors in prediction. Additionally, some models are not suitable for driving environment due to the high computational cost.

On the other hand, very recent works have explored the use of vision-language models for facial expression recognition yielding very interesting results [11], [27], [4], and [13]. For instance, Li *et al.* [11] proposed a Contrastive Language-Image Pretraining(CLIP)-based framework for dynamic and static facial expression recognition, incorporating fine-grained text descriptors for each expression.

(a) Camera installed in a car [3]. **(b) Facial expression recognition for a driver [2].**

Figure 1: Example of driver's facial expression recognition using a camera installed inside a car.

Similarly, Zhao and Patras [27] focused on dynamic facial expression recognition, incorporating temporal modeling and learnable context to capture fine-grained temporal features. Foteinopoulou and Patras [4] addressed zero-shot classification challenges in dynamic facial expression recognition by using sample-level text descriptions for natural language supervision. A quantitative assessment of GPT-4V's performance in General Emotion Recognition (GER) tasks is provided in [13]. These few recent works showed interesting results and proved the usefulness of vision language models in facial expression recognition.

Inspired by the methods above, we propose a novel architecture that exploits the capabilities of the vision-language models for facial expression recognition in a driving environment. In fact, while the existing VLM-based approaches have shown promising results, they do not deal with driving environments and may struggle to achieve high accuracy in real-world scenarios due to their dependence on complex textual descriptors and the high computational costs involved. We propose an elegant approach that uses the Contrastive Language-Image Pretraining [16] and combines both visual and textual data to overcome the environmental challenges and ambiguities in facial expression interpretation in a driving environment. In addition, we apply average pooling to the features extracted by CLIP to reduce the dimensionality and highlight the salient information, thereby reducing the computational cost and improving the performance of the classifier. The contributions of our work are described as follows:

- We introduce CLIVP-FER, an innovative approach for extracting features from both images and text using the CLIP model, along with Multilayer Perceptron (MLP) classifier for accurate classification. This method effectively leverages multimodal data to provide a detailed understanding of the driver's facial expressions.
- Our methodology harnesses the CLIP model solely as a feature extractor, without additional training. This strategy enables us to leverage the strengths of a pre-trained model, significantly reducing computational costs.
- We apply average pooling to the features extracted by CLIP to reduce dimensionality of the features and highlight salient information, thereby improving the performance of the classifier.
- We conduct a comprehensive evaluation of the proposed approach, which includes assessing its generalization capabilities, speed efficiency, and performing an ablation study.
- We compare our approach to the state-of-the-art and existing methods on benchmark dataset of a driving environment, resulting in significant performance improvements.

The rest of this paper is structured as follows: Section 2 provides a review of existing works related to facial expression recognition. Section 3 describes our proposed approach, detailing the different steps from data pre-processing to feature extraction with CLIP to emotion classification. Section 4 discusses the experimental data and setup. Section 5 presents the obtained results and compares them to state-of-the-art. To better gain insights into our proposed architecture, cross-database evaluation, and ablation study are also provided in this section. Finally, Section 6 summarises our findings, draws some conclusions, and suggests possible directions for future research.

2 Related Work

Several studies have investigated the challenging task of identifying driver's emotions, primarily focusing on physiological signals and facial expression cues. In the realm of physiological signals, Wang *et al.* employed electroencephalography (EEG) signals and proposed an end-to-end Convolutional Neural Networks (CNN) model in order to improve the cross-subject emotion recognition accuracy [23]. By doing this, the authors achieved a good performance in determining human emotions. In another work, using a back-propagation network and Dempster–Shafer evidence approach, Wang *et al.* employed multiple-electrocardiogram (ECG) feature fusion including

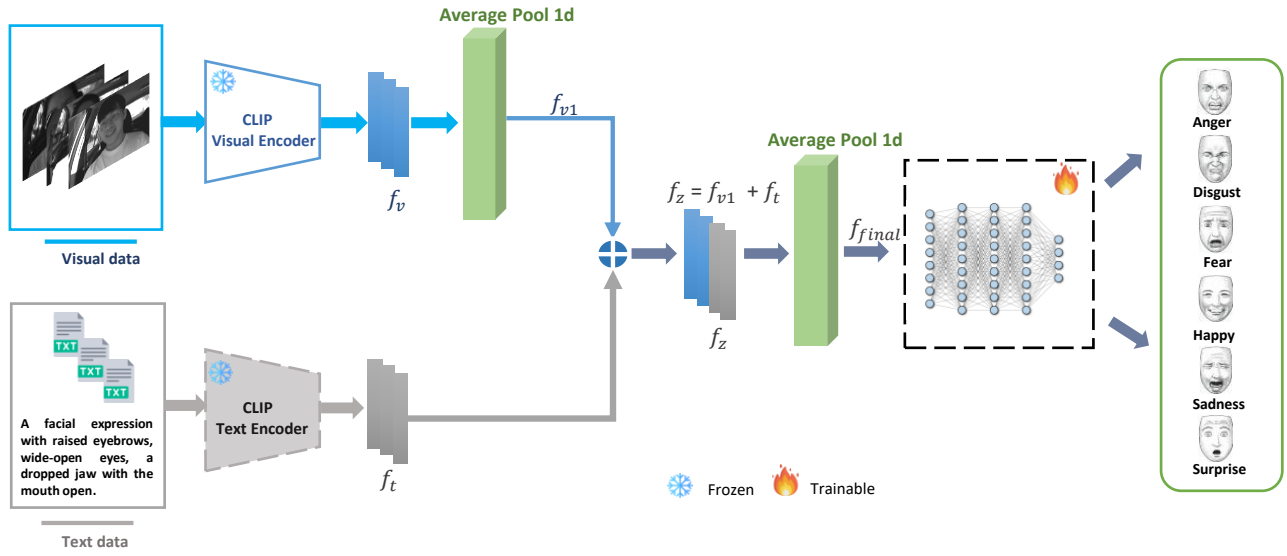

Figure 2: The figure illustrates our proposed CLIVP-FER architecture. The visual inputs are processed with Contrastive Language-Image Pretraining(CLIP) [16] encoder into a vector f_v. The textual inputs are processed into a vector f_t. The visual feature vector f_v is then processed through an average pooling layer to produce f_{v1}, which is subsequently combined with f_t to form f_z. This combined vector is further processed by an additional average pooling layer, resulting in the final feature vector f_{final}, used to classify the facial expressions by a Multilayer Perceptron (MLP). The figure highlights the frozen (non-trainable) and trainable components.

time-frequency domain, waveform, and non-linear features to recognize driver's emotion [22]. In contrast, various other approaches have focused on visual data for facial expression inference. For instance, Chen *et al.* introduced a modified ResNet18 network with an enhanced feature attention (EFA) module to extract rich features from facial expression images [1]. They employed a joint discriminative correlation alignment (JDCA) loss to align feature distributions between single-driver (SD) and two-driver (TD) images while leveraging label information for driver facial expression recognition. In [20], the authors used transfer learning with pre-trained CNN architectures such as AlexNet, SqueezeNet, and VGG19 to perform facial emotion recognition. The models were evaluated on various benchmark datasets with real-time in-vehicle challenges. Results indicated that the pre-trained VGG19 model generally outperforms AlexNet and SqueezeNet.

In [7], a novel facial expression recognition method designed for real-time embedded systems is proposed. The method utilizes a DLib detector [9] to detect the face landmarks and extract geometric features. The extracted features are then employed using a hierarchical Weighted Random Forest classifier to accurately classify the facial expressions, claiming interesting results.

To analyze driver's behavior, a pre-trained VGG16 model was used in [10] to extract features and perform classification of emotions under the challenges of multi-pose and varying illumination conditions, achieving interesting results. In the same context, a real-time framework for stress detection was presented in [25]. It consists of three modules, face detection using MTCNN, a connected convolutional network (CCNN) that combines low-level and high-level features for the facial expression module, and a module for stress detection. The proposed framework achieved a performance comparable to that of a state-of-the-art one.

Among the most recent and appealing works on facial expression recognition are those exploiting vision-language models [11], [27], [4], [13], and [21]. These works showed interesting results and proved the usefulness of vision language models in facial expression recognition. However, their efficiency in real-world driving environments is not yet proven. To the best of our knowledge, our present paper is the first work focusing on exploring the capabilities of the vision-language models for facial expression recognition in driving environments. A comprehensive survey on driver's emotion recognition can be found here [18].

3 Proposed Approach

Our proposed architecture, called CLIVP-FER, is illustrated in Figure 2. The inputs of the system consist of a face image and a tokenized text descriptor. Feature extraction is independently conducted on each input utilizing CLIP [16] encoder. Subsequently, a two-stage average pooling layer (AvgPool1d) is applied to the image features and to the concatenated image-text features, in order to enhance the feature saliency and decrease the dimensionality. These refined features are then fed into a Multilayer Perceptron (MLP) classifier for determining the facial expression of the driver.

3.1 Data Preprocessing

For the visual data, and to enhance the model's robustness and prevent overfitting, a series of transformations are applied to the input images, including random horizontal flip and random rotation, hence introducing variability that simulates different orientations

and perspectives seen in real-world scenarios. Additionally, resizing and normalization are applied to the images.

The textual data in our study consists of descriptive captions for each class, detailing the corresponding facial expressions. These captions are generated by the advanced language model, ChatGPT-4, rather than using standard class names. For example, for the 'Happy' class, the caption is: 'A facial expression characterized by wide, bright eyes, raised cheeks, a broad smile revealing teeth, and relaxed eyebrows.' This strategy, inspired by the work presented in [27], allows for a more detailed and nuanced representation of facial expressions. The preprocessing of these captions involves tokenization and embedding using CLIP to convert the tokens into feature vectors that can be processed alongside the visual data.

3.2 Extracting and Refining Features Using CLIP and Average Pooling

We use CLIP, a multimodal neural network trained on a large number of text-image pairs, to extract the features. CLIP has been designed with two distinct encoders: a visual encoder, E_{visual}, and a text encoder, E_{text}, which operate independently to handle their respective data modalities. For our purposes, we select the pre-trained ViT-B/32 CLIP variant, which employs a Vision Transformer (ViT) architecture. The ViT-B/32 is especially useful since it is able to generate highly expressive and contextual features for both text and images, which is critical for capturing the nuances of facial expressions, especially in scenarios where visual data might be partially occluded or unclear. During feature extraction, CLIP encoders process the preprocessed images and text without additional training. Let V denote the preprocessed image, and T the preprocessed text. The feature extraction is defined as follows:

$$f_{\text{v}} = E_{\text{visual}}(V) \tag{1}$$

$$f_{\text{t}} = E_{\text{text}}(T) \tag{2}$$

where f_{v} and f_{t} are the image and text feature vectors, respectively.

The resulting image feature vector undergoes average pooling, which condenses information, reduces dimensionality, and discards noise:

$$f_{\text{v1}} = \text{AvgPool1d}(f_{\text{v}}) \tag{3}$$

The process of average pooling is directly applied to the high-level features extracted by CLIP's vision transformer-based visual encoder. This contrasts with applying pooling layer to features extracted by CNN and leverages the advanced representations provided by the transformer architecture. Subsequently, a second average pooling operation is performed on the combined image-text features set. This two-stage pooling process ensures focusing on the most relevant features for classification:

$$f_{\text{z}} = \text{Concat}(f_{\text{v1}}, f_{\text{t}}) \tag{4}$$

$$f_{\text{final}} = \text{AvgPool1d}(f_{\text{z}}) \tag{5}$$

Finally, the condensed feature set f_{final} is used to train our classifier.

3.3 Emotion Classification

We utilize a specialized MLP classifier with a hidden layer that is specifically designed for categorizing emotions. This classifier is trained to accurately interpret the extracted set of final features, f_{final}, by CLIP. The structure of the classifier is as follows: The initial fully connected layer, known as `fc1`, transforms the input vector into a hidden representation that consists of 512 dimensions. This transformation is achieved using the equation:

$$h = \text{ReLU}(W_1 f_{\text{final}} + b_1) \tag{6}$$

Here, the weights and bias of `fc1` are denoted as W_1 and b_1, respectively. The Rectified Linear Unit activation function, ReLU, is applied in this process. To enhance generalization and prevent overfitting, a dropout layer is implemented with a rate of 0.5 on this hidden representation. The final layer, referred to as `fc2`, maps the hidden representation to the output space that corresponds to the number of emotion categories. This mapping is achieved using the equation:

$$o = W_2 h + b_2 \tag{7}$$

In this equation, W_2 and b_2 represent the weights and bias of `fc2`, respectively. The output denoted as o, represents the raw scores for each emotion category.

4 Experimental Data and Setup

We carried out a comprehensive evaluation of the proposed architecture using a publicly available benchmark dataset of a driving environment namely: KMU-FED (Keimyung University Facial Expression of Drivers) [7]. Example images from this dataset are shown in Figure 3. Additionally, we considered two other datasets (FER 2013 [5], and RAF-DB [12].) for cross-database analysis to assess the generalization of our approach.

The KMU-FED dataset [7] provides an exceptional context for evaluating the effectiveness of our method in real-life driving scenarios, especially due to its emphasis on driver-specific facial expressions. It comprises 1106 images of 12 subjects, each displaying the six basic emotions: anger, disgust, fear, happiness, sadness, and surprise. The images were captured in real driving conditions using near-infrared cameras. These cameras were subject to varying lighting conditions and partial occlusion. To ensure statistically consistent experimentation, we used a 10-fold cross-validation approach to split the dataset.

We implemented our approach using the open-source PyTorch framework on an NVIDIA Quadro RTX 5000 GPU with 16GB RAM. Facial image pre-processing involved the use of Multi-task Cascaded Convolutional Networks (MTCNN) [26] to detect and crop the faces in the KMU-FED dataset, all images in this dataset were resized to 224×224 pixels. During the training phase, we set a batch size of 64, a learning rate of 0.003, and a weight decay of 1e-4. We employed the Adaptive Moment Estimation (Adam) optimizer, cross-entropy loss function, and trained for 40 epochs. Early stopping was implemented when no accuracy improvement was obtained after 10 epochs.

5 Experimental Results and Analysis

This section describes the obtained results using our proposed approach on the KMU-FED dataset along with a comparative analysis against some state-of-the-art methods. In addition, a cross-database evaluation is given assessing the generalization ability and speed efficiency of our approach. Finally, an ablation study is presented

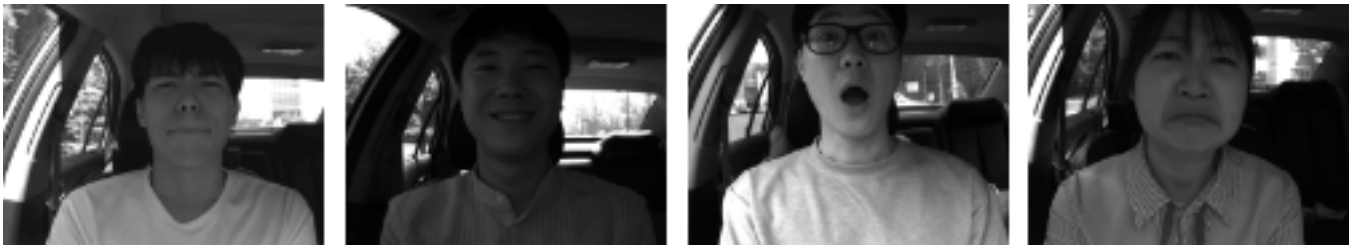

Figure 3: Examples of images from KMU-FED dataset. As can be noticed, the problem of facial expression recognition in driving environment has some different challenges (in terms of lighting and user's cooperation) compared to "conventional" facial expression recognition.

giving more insights into the performance of our proposed archi tecture.

5.1 Obtained Results

Table 1 shows the obtained result using our CLIVP-FER model an a comparison against some recent and state-of-the-art methods o the KMU-FED dataset. The table clearly indicates that our approacl yields an impressive performance of 97.36 % outperforming al other methods. It shows a gain of 2.66 % over the hierarchical WR method, 2.26 % over the LMRF method, and 3.09% over the pre trained VGG16 method. The confusion matrix in Figure 4 confirm the accuracy of our model, with high accuracy for 'Happy' an 'Surprised', and some confusion between 'Angry' and 'Sad', as wel as between 'Fear' and 'Sadness', which can be attributed to th subtle visual similarities of these emotional states. Despite thes few cases of misclassification, the result confirms CLIVP-FER' ability to accurately interpret complex emotions, demonstrating it potential for real-world applications.

Table 1: Obtained results using our CLIVP-FER model, compared with state-of-the-art methods on the KMU-FED dataset.

Methods	Accuracy
Hierarchical WRF (2018) [7]	94.70%
LMRF (2020) [8]	95.10%
Pre-trained-VGG16 (2021) [10]	94.27%
Modified SqueezeNet (2022) [19]	83.40%
CLIVP-FER (Our method)	**97.36%**

5.2 Cross-Database Evaluation

In order to thoroughly investigate the generalisability and speed efficiency of our proposed CLIVP-FER approach, we performed a cross-database analysis by training our model on the KMU-FED dataset and evaluating it on the FER2013 and RAF-DB datasets.

As shown in Table 2, various performance metrics, including accuracy, precision, recall and F1 score, are reported to measure the model's ability to correctly identify facial expressions. Precision and recall, although closely related to accuracy, provide more detailed information about the performance of the model. Precision indicates how many of the positive outcomes predicted by the

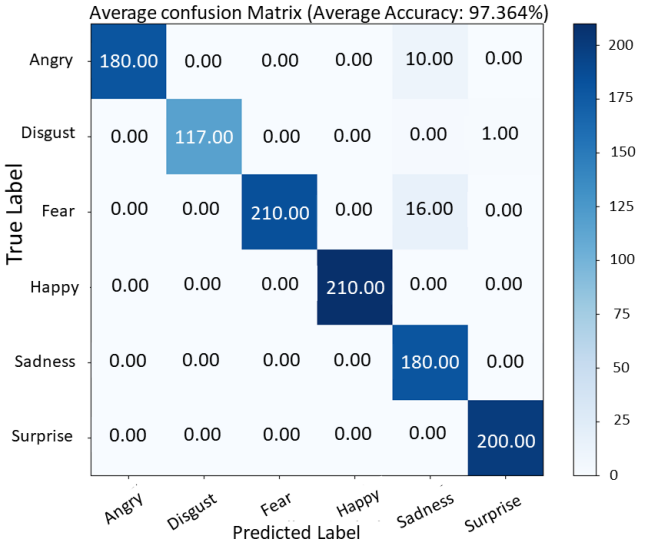

Figure 4: Confusion matrix of our CLIVP-FER model on the KMU-FED dataset.

model were actually positive, and recall indicates how many of the actual positive outcomes were correctly predicted by the model. The F1 score, a combination of precision and recall, indicates a well-balanced model that is both accurate and sensitive.

When trained on the KMU-FED dataset (i.e. which corresponds to a driving environment), the model achieves an accuracy of 0.89 on FER2013 (non-driving environment) and 0.92 on RAF-DB (non-driving environment). For a fair comparison, given the fact that our model was trained on 6 classes, we tested the model on the same number of classes, excluding the seventh class (Neutral) from both test datasets. The model is still shown to be quite accurate, with a slight drop in performance when tested on unbalanced samples, which is due to the balanced nature of its training set. This decrease in accuracy can also be attributed to the fact that certain features present in the FER2013 and RAF-DB datasets are somehow different from those in driving environment, such as those found in the KMU-FED dataset.

To evaluate the speed efficiency of our model, we recorded the inference time across the two datasets, we achieved an average

inference time of 2.6 ms per image on the FER2013 dataset and 2.7 ms on the RAF-DB dataset. These results not only demonstrate the robustness of our model across the two emotional datasets but also highlight its rapid inference capabilities, which are essential for real-time applications.

In summary, our CLIVP-FER model has shown interesting generalization capabilities and efficient inference times, even though it is trained solely on a driving environment dataset and tested on non-driving environment (i.e. the FER2013 and RAF-DB datasets). Its ability to maintain high accuracy across different datasets and achieve rapid inference times indicates its robustness and potential for practical applications, particularly in driving contexts where accurate facial expression recognition is crucial.

Table 2: Cross-database performance evaluation of our proposed CLIVP-FER model. KMU-FED dataset corresponds to a driving environment while FER2013 and RAF-DB datasets correspond to a non-driving environment. The training is conducted on KMU-FED and the testing is on FER2013 and RAF-DB datasets.

Datasets	KMU-FED	
	FER2013 [5]	RAF-DB [12]
Accuracy	0.89	0.92
Precision	0.92	0.85
Recall	0.87	0.87
F1-score	0.88	0.85

5.3 Ablation Study

In order to determine the impact of different image- and text-derived features on the overall performance of our CLIVP-FER model, an ablation study is conducted. The results of this ablation study are detailed in Table 3. Using only CLIP image features, the model yields in quite low performance, achieving an average accuracy of 0.63, a precision of 0.62, a recall of 0.60 and an F1 score of 0.62. Incorporating textual features in addition to the image features significantly improves the model's performance, as evidenced by a better average accuracy of 0.84, highlighting the benefits of the fusion (image and text). The gains in precision (0.88) and recall (0.81) further highlight the model's refined ability to accurately classify expressions and reliably identify the majority of classes. Our proposed approach, taking advantage of a concatenated set of image and text features combined with average pooling layer, achieves excellent performance (accuracy: 0.97, precision: 0.98, recall: 0.97 and F1 score: 0.97). These results highlight the importance of integrating multimodal features into the model and demonstrate the importance of each component of the system.

6 Conclusion and Future Work

In this paper, we introduced CLIPV-FER, a novel approach that utilizes the CLIP vision-language model to recognize facial expressions in driving scenarios. Our approach concatenates the features extracted by a pre-trained CLIP variant from images and text descriptions, leading to a more comprehensive representation of facial expression recognition. The feature set is further refined by applying average pooling, enabling the classifier to train with more focus and pertinent information while minimizing the dimensionality of the features. We assessed the performance of our approach on a benchmarking dataset of a driving environment and conducted a cross-database evaluation that demonstrated good generalization ability and speed efficiency of our approach. The experimental results showed significant performance enhancement, highlighting the effectiveness of our approach in recognizing facial expressions in a driving environment.

This work is by no means complete. First, the findings should be further validated using other driving datasets, despite the current scarcity of such datasets. Another direction for future research is to investigate how the model performs with video data, which would capture the temporal changes in facial expressions. Additionally, exploring the integration of other modalities such as data from wearable sensors, could enhance the model's performance. Of interest is also the exploration of a multi-camera system in the driving scenario, which would be beneficial for capturing and handling various head poses.

Table 3: Evaluating the impact of CLIVP-FER model's components on the KMU-FED dataset. The first row shows the results of using only image features. The second row shows the results of combining image and text features. The last row demonstrates the performance of our model utilizing both visual and textual features, along with an average pooling layer.

Methods	Accuracy	Precision	Recall	F1-score
Visual	0.63	0.62	0.60	0.62
Visual+Text	0.84	0.88	0.81	0.81
Visual+Text+AvgPool	**0.97**	**0.98**	**0.97**	**0.97**

7 Acknowledgments

We wish to convey our deep appreciation to the EUNICE Alliance for the financial support. Abdenour Hadid is funded by TotalEnergies collaboration agreement with Sorbonne University Abu Dhabi.

References

[1] Xiaobo Chen, Jian Du, Fuwen Deng, and Feng Zhao. 2023. Transferable driver facial expression recognition based on joint discriminative correlation alignment network with enhanced feature attention. *IET Intelligent Transport Systems* 17, 12 (2023), 2444–2457.

[2] EPFL. 2014. Face analysis for automotive applications. https://www.epfl.ch/innovation/domains/transportation/vehicles/intelligent-vehicles/face-analysis-for-automotive-applications/ Accessed: 29 April , 2024.

[3] Sebastian Fillenberg. 2016. Continental bringt Biometrie ins Fahrzeug. https://www.continental.com/de/presse/pressemitteilungen/2016-12-15-biometrics/ Accessed: 29 April , 2024.

[4] Niki Maria Foteinopoulou and Ioannis Patras. 2023. EmoCLIP: A Vision-Language Method for Zero-Shot Video Facial Expression Recognition. *arXiv preprint arXiv:2310.16640* (2023).

[5] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*. Springer, Daegu, Korea, 117–124.

[6] Deepak Kumar Jain, Ashit Kumar Dutta, Elena Verdú, Shtwai Alsubai, and Abdul Rahaman Wahab Sait. 2023. An automated hyperparameter tuned deep learning model enabled facial emotion recognition for autonomous vehicle drivers. *Image and Vision Computing* 133 (2023), 104659.

[7] Mira Jeong and Byoung Chul Ko. 2018. Driver's facial expression recognition in real-time for safe driving. *Sensors* 18, 12 (2018), 4270.

[8] Mira Jeong, Jaeyeal Nam, and Byoung Chul Ko. 2020. Lightweight multilayer random forests for monitoring driver emotional status. *Ieee Access* 8 (2020), 60344–60354.

[9] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research* 10 (2009), 1755–1758.

[10] Alessandro Leone, Andrea Caroppo, Andrea Manni, and Pietro Siciliano. 2021. Vision-based road rage detection framework in automotive safety applications. *Sensors* 21, 9 (2021), 2942.

[11] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. 2023. CLIPER: A Unified Vision-Language Framework for In-the-Wild Facial Expression Recognition. *arXiv preprint arXiv:2303.00193* (2023).

[12] Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2852–2861.

[13] Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. 2024. GPT-4V with emotion: A zero-shot benchmark for Generalized Emotion Recognition. *Information Fusion* 108 (2024), 102367.

[14] MARK PARADIES. 2022. Is Human Error the Cause of 94% of Vehicle Accidents? Would Automation Stop These Crashes? https://www.taproot.com/is-human-error-the-cause-of-vehicle-accidents/ Accessed: 04 December , 2023.

[15] Mrinalini Patil and S Veni. 2019. Driver emotion recognition for enhancement of human machine interface in vehicles. In *2019 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 0420–0424.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[17] Ibtissam Saadi, Douglas W. Cunningham, Taleb-Ahmed Abdelmalik, Abdenour Hadid, and Yassin El Hillali. 2023. Driver's Facial Expression Recognition Using Global Context Vision Transformer. In *2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*. IEEE, Gwalior, India, 1–8. https://doi.org/10.1109/CVMI59935.2023.10464794

[18] Ibtissam Saadi, Douglas W Cunningham, Abdelmalik Taleb-Ahmed, Abdenour Hadid, and Yassin El Hillali. 2024. Driver's facial expression recognition: A comprehensive survey. *Expert Systems with Applications* 242 (2024), 122784.

[19] Goutam Kumar Sahoo, Santos Kumar Das, and Poonam Singh. 2022. Deep Learning-Based Facial Emotion Recognition for Driver Healthcare. In *2022 National Conference on Communications (NCC)*. IEEE, Mumbai, India, 154–159.

[20] Goutam Kumar Sahoo, Santos Kumar Das, and Poonam Singh. 2023. Performance Comparison of Facial Emotion Recognition: A Transfer Learning-Based Driver Assistance Framework for In-Vehicle Applications. *Circuits, Systems, and Signal Processing* (2023), 1–28.

[21] Zeng Tao, Yan Wang, Junxiong Lin, Haoran Wang, Xinji Mai, Jiawen Yu, Xuan Tong, Ziheng Zhou, Shaoqi Yan, Qing Zhao, et al. 2024. A3lign-DFER: Pioneering Comprehensive Dynamic Affective Alignment for Dynamic Facial Expression Recognition with CLIP. *arXiv preprint arXiv:2403.04294* (2024).

[22] Xiaoyuan Wang, Yongqing Guo, Jeff Ban, Qing Xu, Chenglin Bai, and Shanliang Liu. 2020. Driver emotion recognition of multiple-ECG feature fusion based on BP network and D–S evidence. *IET Intelligent Transport Systems* 14, 8 (2020), 815–824.

[23] Zhirong Wang, Ming Chen, and Guofu Feng. 2023. Study on Driver Cross-Subject Emotion Recognition Based on Raw Multi-Channels EEG Data. *Electronics* 12, 11 (2023), 2359.

[24] Khalid Zaman, Zhaoyun Sun, Sayyed Mudassar Shah, Muhammad Shoaib, Lili Pei, and Altaf Hussain. 2022. Driver Emotions Recognition Based on Improved Faster R-CNN and Neural Architectural Search Network. *Symmetry* 14, 4 (2022), 687.

[25] Jin Zhang, Xue Mei, Huan Liu, Shenqiang Yuan, and Tiancheng Qian. 2019. Detecting negative emotional stress based on facial expression in real time. In *2019 IEEE 4th international conference on signal and image processing (ICSIP)*. IEEE, 430–434.

[26] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.

[27] Zengqun Zhao and Ioannis Patras. 2023. Prompting visual-language models for dynamic facial expression recognition. *arXiv preprint arXiv:2308.13382* (2023).