MULTI-MODAL FUSION ENHANCED MODEL FOR DRIVER'S FACIAL EXPRESSION RECOGNITION

Jianrong Chen[*], Sujit Dey[*], Lei Wang[†], Ning Bi[†] and Peng Liu[†]

[*]Department of Electrical and Computer Engineering, University of California, San Diego
[†]Qualcomm AI Research
{jic497, dey}@eng.ucsd.edu, {wlei, nbi, peli}@qti.qualcomm.com

ABSTRACT

Facial expression recognition (FER) for monitoring a driver's emotional state has become an increasing need for advanced driver assistant systems (ADAS). Though state-of-art results of recognition accuracy have been achieved in FER with the development of deep neural networks (DNNs) in recent years, FER in real-world is still challenging due to illumination and head pose variation. In this work, we propose a multi-modal fusion based FER model capable of recognizing facial expressions accurately regardless of the lighting conditions and head poses, using a structured-light imaging camera which provides three modalities of images - RGB, Near-infrared (NIR), and Depth Maps. The model is implemented in two phases, where the first phase extracts feature from single modalities separately using 3D ResNet while the second phase combines the multi-modal features and classifies expressions. The model is trained and tested with a novel facial expression dataset with the three image modalities, with varying lighting conditions and head poses. The experimental results show that combining different modalities improves the model performance and robustness. A recognition accuracy of over 90% has been obtained in the usage scenario of FER for drivers.

Index Terms— Facial expression recognition, multi-modal fusion, deep neural networks, driver emotion monitoring

1. INTRODUCTION

Facial expression recognition (FER) has been widely studied in recent years. It can be applied to various usage scenarios such as human-computer interaction, medical treatment [1], and advanced driver assistant systems (ADAS) [2][3].

FER can be critical for ADAS in improving road safety. Since facial expressions can reflect driver's fatigue condition and emotional state, FER can be very useful for ADAS to recognize the driver's state to provide timely alerts to the driver as well as to neighboring drivers. We also believe that FER can be useful for level 3-4 autonomous driving, since detecting driver's state enables safe control switching between the driver and the vehicle.

Recently, there have been many studies in computer vision for FER system. FER is a multi-class classification task, where typically seven basic emotional expressions defined by Ekman [4] (anger, disgust, fear, happiness, neutral, sadness, and surprise) are to be recognized. There are many publicly available datasets collected based on these basic expressions, such as the extended Cohn-Kanade (CK+) [5], the Oulu-CASIA dataset [6], JAFFE [7], and CMU Multi-PIE [8]. To our best knowledge, most of the datasets only contain data of a fixed head pose collected in a good illumination condition. However, in real-world in-vehicle driving conditions, the illumination condition is not always good and the driver's head pose varies, which may affect the accuracy of facial features extraction and thus lead to an inferior performance of the FER model when applied to real-world situations. Hence, to tackle these challenges, *we develop a novel facial expression dataset with three modalities of images collected simultaneously, i.e., RGB images, Near-infrared (NIR) images, and Depth Maps, with different illumination conditions.* Among these modalities, NIR images and Depth Maps are not affected by the ambient illumination conditions. The images are collected under three different head poses.

It has been shown using widely evaluated benchmarks, such as CK+ and MMI [9], that training networks on image sequences can improve the performance. [10] *In our work, we propose a video-based multi-modal fusion model, which gives better and more robust FER accuracy for different driving conditions.* The model is implemented in two phases. In the first phase, networks based on 3D CNN are trained on RGB, NIR and Depth Map videos separately to extract features from each modality. In the second phase, features of different modalities are combined and input to a multilabel classifier to recognize expressions.

In the remainder of this paper, Section 2 describes the related work. In Section 3, we give an overview of the dataset and describe the data collection and pre-processing steps. In Section 4, we explain our proposed multi-modal fusion based model. Experimental results are given in Section 5, and future work is described in Section 6.

2. RELATED WORK

Deep learning has been increasingly popular in computer vision research and has achieved state-of-the-art performances in FER using datasets such as CK+ and JAFFE. However, the above datasets are only collected in a laboratory environment with good illumination. Models may fail when it comes to real-world driving conditions with bad illumination.

To address problems caused by illumination changes for the FER system, Jeong et. al [3] collected KMU-FED dataset for driver's FER specifically, where NIR images of seven basic expressions are collected in a vehicle. Zhao et al. [6] collected the Oulu-CASIA dataset containing both NIR and RGB images and proved that more robust FER results against illumination variations can be obtained by using NIR images instead of using RGB images under poor illumination with illumination enhancement. However, the KMU-FED dataset only has NIR images, and the RGB and NIR images in the Oulu-CASIA dataset are not synchronized, which makes it unlikely to develop a model with better performance by fusing abundant information from different image modalities in these datasets. Besides, all the above datasets only have images captured from the frontal view, while in real-world a driver can have various head poses relative to the camera.

978-1-6654-4989-2/21/$31.00 ©2021 IEEE

Therefore, we construct a novel facial expression dataset with RGB images, NIR images, and Depth Maps collected simultaneously, consisting of data collected with three different head poses. We propose multi-modal fusion based model to recognize facial expressions accurately, robust to the illumination conditions and head poses.

3. DATA COLLECTION

We collect data of facial expressions from 20 subjects with 3 head poses. Besides seven basic expressions, the data also includes the yawning expression, which is necessary for fatigue surveillance for drivers. Images of three modalities (RGB images, NIR images, and Depth Maps) are collected simultaneously. Data pre-processing and augmentation is performed before feeding the data into the model.

3.1 Dataset summary

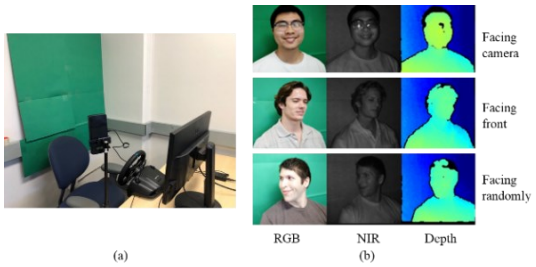

Fig. 1. (a) Facial expression data collection set up and (b) an example of collected images

The data collection is conducted in a laboratory setup shown in Fig. 1(a). A structured-light camera device is used to capture images of the subject's upper body, including RGB images (1080x1920 pixels), NIR images (184x324 pixels) and Depth Maps (124x216 pixels), which are synchronized and collected simultaneously. The Depth Map reflects the distance between the subject to the camera. The camera is mounted around the rearview mirror position relative to the subject, with a distance of around 0.7 meter to the subject. During data collection, the subjects are asked to imitate and make facial expressions corresponding to certain emotions. Videos of eight types of facial expressions are recorded several times with the subject facing the front, the camera, and a random direction. Each facial expression sequence is manually annotated and extracted from the raw data. Examples of the collected images are shown in Fig.1.(b). Table 1 is a summary of the dataset. More example videos are included in the supplemental material.

3.2. Data pre-processing

We pre-process the raw data by (1) cleaning the dataset and (2) detecting and extracting the face image. We manually checked every clip to delete extreme samples where the participant's expression is either too mild or ambiguous and also excluded contents without the target expression in each clip, especially at the beginning/end of the clip where the facial expression is not posed yet.

To extract the most useful information from the raw data, face extraction is also required. For RGB images, we implement a face normalization algorithm [11] by detecting and aligning face landmarks to normalize the face shown as Fig.2.(a). Compared with extracting the face directly from the bounding box provided by a face detector, by face aligning and cropping, we can exclude the noise introduced by head movements.

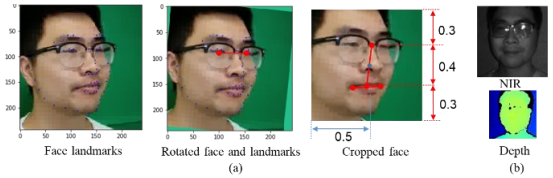

Fig. 2. (a)Face alignment and extraction for RGB images (b) Face extraction for NIR and Depth Map

Table 1. Dataset summary

Subject	Number	20 (8 female)
	Age	20-55 years
Expression (number of sequences)	Anger(268), Disgust(212), Fear(116), Happiness(299), Neutral(336), Sadness(346), Surprise(191), Yawning(251)	
Data Sequence	Modality	RGB/NIR/Depth
	Duration	4sec - 10sec
Frame Rate	RGB	30 fps
	NIR/Depth	15 fps
Head Pose	Front, Camera, A Random Direction	

However, for Depth Map, it is difficult to detect the landmarks accurately because of their low resolution. Instead, we detect and align faces from NIR images. Based on the cropping bounding box obtained from the NIR image, we can extract the face part from its corresponding Depth Map, shown in Fig.2.(b). All the cropped face images are resized to 112x112.

3.3. Data augmentation

Since we propose to perform FER based on consecutive frames, the above collected images are augmented by window slicing subsequences of consecutive frames. We implemented a temporal data augmentation method shown in Fig.3.(a). For RGB data, we split a video into several windows of 30-frame clips continuously with 15 frames overlapping. Given the frame rate of NIR and Depth Map is around half of RGB's frame rate, to make the input clips synchronized among different modalities, we extract 16-frame clips continuously with 8 frames overlapping from NIR and Depth Map videos. There are around 1900 data samples for each modality after the temporal augmentation.

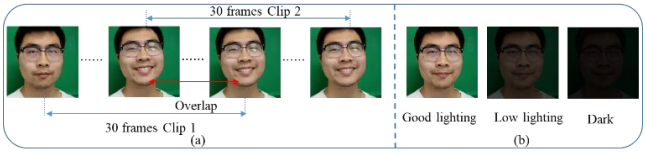

Fig. 3. (a) Temporal data augmentation (b) Lighting emulation

The original data was collected in a laboratory environment with good illumination. Considering our goal is detecting driver's facial expressions in real-world environments with different lighting and shadow conditions, due to the difficulties to collect more data with different illuminations either in a laboratory or real-world environment during the pandemic, we emulate different lighting conditions in the current dataset (collected before the pandemic related restrictions) before developing models. As shown in Fig.3.(b), for each clip, we randomly emulate one of the three lighting conditions by generating a dark mask and applying the mask to the image with a transparency rate. Note that since the NIR and Depth Map are not influenced by ambient lighting conditions, the lighting emulation is only done on RGB images. We train our proposed model using the pre-processed and augmented data.

4. PROPOSED MODEL

In this section, we introduce a robust multi-modal fusion model. With the pre-processed and augmented data, we fine-tune a pre-trained 3D CNN as a feature extractor using consecutive frames as input for each modality. Based on the features extracted from different image modalities by the backbone network, a multilabel classifier is trained to classify facial expressions.

4.1. Backbone network training

In this subsection, we describe the backbone network structure, together with details of the training process. As we stated in Section 2, taking a temporal window of consecutive frames as input to train the network has been shown to give better performance in the context of FER [10]. For this work, 3D ResNet is used as the backbone network to extract features from image sequences, which consists of four 3D residual layers followed by an average pooling layer and a fully-connect layer [12].

To better represent features using 3D ResNet, we utilize a transfer learning technique, i.e., fine-tuning, which is an efficient method widely applied to training tasks on small datasets. In this work, the 3D ResNet is pre-trained on the UCF-101 dataset [13] for action recognition in videos. We fine-tune the network by freezing the first three layers during training. The data is divided into 4 folds for person-independent cross-validation experiments, that is, validate data of three randomly selected subjects and train on the rest of the data. We set up two kinds of training tasks to investigate model performance in general use case and driving-related use case. In the general case, data of all 8 expressions is used for training and validation. For the driving-related use case, considering the most relevant emotions/expressions for a driver that need to be detected, only data of "Neutral", "Anger", "Happiness" and "Yawning" is used. For each case, we train three backbone networks called 3D ResNet-RGB, 3D ResNet-NIR and 3D ResNet-Depth using the three image modalities respectively, which enable effective feature extraction from each modality.

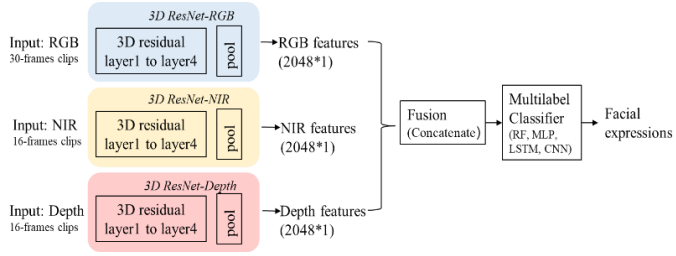

Fig. 4. Multi-modal fusion model structure

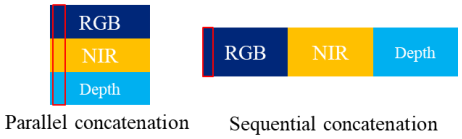

Fig. 5. Concatenation of multi-modal features

4.2. Multi-modal fusion model

In this subsection, we introduce the overall structure of the multi-modal fusion model, where features from different modalities are combined to increase the robustness of the model under various illumination conditions. In this work, we propose an ensemble approach based on feature-level multi-modal fusion. The overall model is shown in Fig. 4. Features are extracted separately by the finetuned backbone networks (3D ResNet-RGB, 3D ResNet-NIR and 3D ResNet-Depth) from three modalities and fused. The features are then concatenated and fed to a multilabel classifier. We experiment with two ways of concatenating features, namely, sequential concatenation and parallel concatenation as shown in Fig. 5, i.e., adding one feature vector after another or parallelly. The backbone networks and the classifier are trained separately.

For this work, we use four kinds of multilabel classifiers: Random Forest (RF) [14], Multi-Layer Perceptron (MLP) [15], Long Short-Term Memory (LSTM) network [16], and CNN classifier. Specifically, we first extract the features of each modality from its corresponding backbone network, then we train the concatenated features on four kinds of classifiers to classify facial expressions.

5. EXPERIMENTAL RESULTS

In this section, we first provide results of the backbone networks trained on each single modality. Then we present an overview of results of our multi-modal fusion model where features from different modalities are combined and trained.

The results are presented in the form of facial expression recognition accuracy, when considering all the eight expressions (General Case) and the driving related expressions, "Neutral", "Anger", "Happiness" and "Yawning" (Driving Case).

5.1. Single modality analysis

In this subsection, we analyze the recognition results of the backbone networks, which are trained on data of a single modality. We train the network on the RGB data with/without lighting emulation in both training tasks; the results are shown in Table 2. The overall recognition accuracy achieved in the original RGB data proves that the 3D ResNet can well represent facial expression features. However, the performance is worse on more realistic data which is emulated with different illumination conditions.

We also finetune the 3D ResNet on the NIR and Depth data. The results of the 3D ResNet-NIR and 3D ResNet-Depth are shown in Table 2. The recognition accuracy in NIR and Depth Map data is lower than that in RGB data without lighting emulation since they do not have as much information due to lower image resolution and frame rate. However, the 3D ResNet-NIR achieves 86.8% accuracy in the driving case, which is higher than that of 3D ResNet-RGB (82.6%) when considering different illumination conditions.

5.2. Multi-modal fusion

Next, we will present results of the fusion of different modalities. As stated in Section 4, the concatenated features are fed into four kinds of classifiers, among which, the RF model and the MLP model only accept 1-D input vectors, thus only sequential concatenation is experimented on them. The LSTM network consists of two LSTM layers with 128 cells each. The CNN classifier consists of one 1-d convolutional layer followed by a ReLU activation layer, a 1-d max-pooling layer and a fully-connect layer. Both concatenation methods are experimented on the LSTM network and the CNN classifier. The metrics for the individual modalities and comparison of different classifiers with the fusion of modalities are presented in Table 2.

As can be seen in Table 2, by fusing different modalities, we can get much higher recognition accuracy than just using RGB-only or any one single modality. For example, while 58.9% and 82.6/86.8% accuracy can be achieved using RGB-only and RGB/NIR-only modality in the general and driving cases respectively, using all 3 modalities and the CNN

Table 2. Results of single-modal backbone network and multi-modal fusion classifiers.
LE: Lighting Emulation, (s): sequential concatenated features input, (p) parallel concatenated features input

General case (RGB(w/o LE) only: 65.3%, RGB(w/LE) only: 58.9%, NIR only: 58.2%, Depth only: 35.0%)						
Modality \ Classifier	RF	MLP	LSTM (s)	LSTM (p)	CNN (s)	CNN (p)
RGB+NIR	**66.4%**	64.6%	49.5%	61.2%	64.7%	65.0%
RGB+NIR+Depth	66.4%	68.1%	40.4%	61.9%	67.4%	**68.9%**
Driving case (RGB(w/o LE) only: 91.3%, RGB(w/LE) only: 82.6%, NIR only: 86.8%, Depth only: 53.1%)						
Modality \ Classifier	RF	MLP	LSTM (s)	LSTM (p)	CNN (s)	CNN (p)
RGB+NIR	89.1%	87.9%	87.8%	90.4%	89.3%	**91.0%**
RGB+NIR+Depth	89.3%	89.3%	60.1%	89.3%	89.1%	**90.7%**

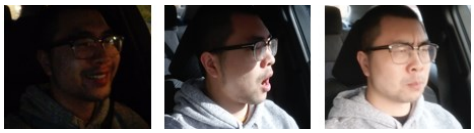

Fig. 6. Example images of driving test data

classifier trained on parallel concatenated features, CNN(p), improves accuracy to 68.9% and 90.7% for the general and driving cases respectively.

While the experimental results show that our proposed multi-modal fusion model can achieve a high accuracy of over 90% in detecting driver's facial expressions, we need to ensure the model is also robust to driver's head pose variation. We achieved 91.7%, 90.5% and 89.1% accuracy under facing camera, front and random directions using the CNN classifier trained on parallel concatenated features of three modalities. The results show that good model performance is maintained across different head poses, even when the driver face has random directions.

Test with Driving Data: To further demonstrate the effectiveness of our proposed multi-modal fusion model in real-world driving conditions, we also collected data of driving related expressions from two subjects inside a vehicle with different realistic lighting and shadow conditions and head poses. The real-world test data contains 72 data samples. Fig. 6 shows a few example driving scenarios, with more images and videos included in the supplemental material. While our single modality 3D ResNet-RGB model could obtain an FER accuracy of 80.6%, our multi-modal fusion model improves FER accuracy to 90.3% on the real-world test data. The results show that good model performance is maintained even in real-world test cases.

6. CONCLUSIONS AND FUTURE WORK

This work proposed a novel multi-modal fusion model for facial expression recognition based on image sequences of RGB, NIR and Depth Map. The ensemble of the 3D ResNet and a multilabel classifier is implemented in the framework. A novel facial expression dataset consisting of images of RGB, NIR and Depth Map is created and augmented with realistic lighting conditions and head poses reflecting driving scenarios. The results demonstrate that significant advantages in both recognition accuracy as well as robustness with regards to lighting conditions and head poses can be achieved using multiple modalities compared to a single modality.

In the future, we plan to augment our dataset by collecting more real-world driver data with our V2X enabled vehicle, following our recently approved IRB protocol, and make it publicly available for research use. We also plan to extend the proposed model to derive driver's state of mind (SoM), where other contributors of SoM (e.g., distraction, fatigue, anxiety) will also be detected.

7. ACKNOWLEDGMENTS

This work is funded by Qualcomm. The authors extend their appreciation to Chienchung Chang and Zhen Wang at Qualcomm, and Professor Truong Nguyen at UCSD, for valuable discussions and feedback.

8. REFERENCES

[1] M. I. U. Haque and D. Valles, "A Facial Expression Recognition Approach Using DCNN for Autistic Children to Identify Emotions," *2018 IEMCON*, Vancouver, BC, 2018, pp. 546-551.

[2] M. A. Assari and M. Rahmati, "Driver drowsiness detection using face expression recognition," 2011 ICSIPA, Kuala Lumpur, 2011, pp. 337-341.

[3] M. Jeong and B. C. Ko, "Driver's Facial Expression Recognition in Real-Time for Safe Driving," Sensors, vol. 18, no. 12, p. 4270, 2018.

[4] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." Journal of personality and social psychology, vol. 17, no. 2, pp. 124–129, 1971.

[5] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 CVPR - Workshops, San Francisco, CA, 2010, pp. 94-101.

[6] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M.PietikäInen, "Facial expression recognition from near-infrared videos". Image and Vision Computing, August 2011, vol. 29, pp. 607-619.

[7] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on. IEEE, 1998, pp. 200–205.

[8] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," Image and Vision Computing, vol. 28, no. 5, pp. 807–813, 2010.

[9] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, 2010, p. 65.

[10] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," in *IEEE Transactions on Affective Computing*.

[11] D. L. Baggio, Mastering OpenCV with Practical Computer Vision Projects. Birmingham, U.K.: Packt Publishing, 2012.

[12] K. Hara, H. Kataoka and Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," in CVPR 2018, Proceedings, IEEE conference, 2018, pp. 6546-6555.

[13] K. Soomro, A. R. Zamir, and M.Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, December 2012.

[14] M. Pal, "Random forest classifier for remote sensing classification," Int. J. Remote Sens., vol. 26, no. 1, pp. 217–222, 2005.

[15] M. W. Garner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences," Atmosph. Environ., vol. 32, no. 14/15, pp. 2627–2636, Aug. 1998.

[16] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in Proc. 9th Int. Conf. Artificial Neural Networks, pp. 850–855, 1999.