Development of Video-Based Emotion Recognition System using Transfer Learning

Teddy Surya Gunawan
Electrical and Computer Engineering Dept.
International Islamic University Malaysia
53100 Kuala Lumpur, Malaysia
tsgunawan@iium.edu.my

Muhammad Nuruddin Muktaruddin
Electrical and Computer Engineering Dept.
International Islamic University Malaysia
53100 Kuala Lumpur, Malaysia
nuruddinamokh@gmail.com

Mira Kartiwi
Information Systems Department
International Islamic University Malaysia
53100 Kuala Lumpur, Malaysia
mira@iium.edu.my

Yasser Asrul Ahmad
Electrical and Computer Engineering Dept.
International Islamic University Malaysia
53100 Kuala Lumpur, Malaysia
yasser@iium.edu.my

Tina Dewi Rosahdi
Department of Chemistry
UIN Sunan Gunung Djati
Bandung, Indonesia 40614
tina_dr@uinsgd.ac.id

Ulfiah
Department of Psychology
UIN Sunan Gunung Djati
Bandung, Indonesia 40614
ulfiah@uinsgd.ac.id

2022 IEEE 8th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA) | 978-1-6654-8800-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSIMA55652.2022.9928867

***Abstract*— Due to the complexity of its system and the numerous advantages of its implementation, video emotion identification is a popular area of study today. There have been multiple ways implemented. In this project, emotion recognition in videos will be performed using deep learning. The model will include initialization, feature extraction, emotion categorization, and prediction. LeNet and AlexNet, two distinct neural networks, are used to extract features and classify emotions. There will be some parameter tuning to determine if it enhances the employed architecture's performance, including optimizers and batch size. Each architecture for deep learning will generate a final forecast of four fundamental emotions: disgust, happiness, sadness, and surprise. AlexNet's performance is enhanced by the SGD optimizer, whereas RMSprop improves LeNet's performance. Results showed that AlexNet with SGD optimizer provides 93.00% recognition accuracy.**

Keywords—emotion recognition, transfer learning, AlexNet, LeNet, accuracy.

I. INTRODUCTION

Emotions are responses to particular circumstances or conditions. Emotions are essential to life as a smile indicates happiness, frowns when angry, and cries when sad. It can be seen that emotion plays a very crucial role in human communication. Human communication will be more efficient if the emotion is understood well. Thus, many researchers have started to develop emotion recognition. Emotions can be recognized through facial and speech literally. There are two types of facial emotion recognition: static (image) and dynamic (video).

The growth of technology gives a lot of benefits to emotion recognition studies as the implementation of artificial intelligence in systems, for instance, deep learning. Over the year, video emotion recognition accuracy keeps improving, showing a good sign to be fully applied in real life. For example, Disney, one of the big entertainment companies, uses video emotion recognition to track the facial expressions of an audience in the theatre so they can know which scene of the movie will impact the viewer. Thus, the next production of movies can be improved. As the implementation of video emotion recognition gives many benefits, it motivated more research to be done so it can be a perfect system.

Deep CNN architectures were utilized in [1, 2]. Specifically, AlexNet and VGG-CNN-M-2048 CNN types are employed. OpenCV's Viola & Jones was utilized for face detection. It is fine-tuning the system using a combination of datasets and EmotiW training. CNN was also used in [3, 4], albeit with distinct architectures, namely VGG-Net, ResNet-50, and DenseNet-121. The VGG-Face and ResNet-50 frameworks are being intensively supervised, increasing validation precision. The experiment is conducted by combining each architecture type and additionally incorporating FG-Net. In [5], the CNN-RNN and C3D hybrid architecture was utilized, which produces better recognition accuracy.

Four processes are involved in facial emotion recognition (FER). First, image input is required. Next, picture preprocessing distinguishes landmarks and facial components from the face area. Then, extraction of temporal and spatial face components and landmark features. Expression classification generates the output for the input image based on the extracted characteristics. There are two distinct forms of FER: static and dynamic. Static FER is dependent solely on fixed facial and frame features. While dynamic FER, also known as video-based FER, has variable facial and framing parameters. For dynamic FER, different procedures will be implemented. Before image processing, the video will be segmented into picture series based on scenarios and shots. It will be divided into frames that will be stored in a specific location.

There are many images and video databases available for emotion recognition [6]. Indian Spontaneous Expression Database (ISED) [7] would be chosen as the database for this study. The input video will be extracted into frames, which will undergo alignment and facial recognition processing. Using transfer learning will be the input for two deep neural networks, LeNet and AlexNet. In deep neural networks, feature extraction and emotion classification serve as objectives. Significantly, both networks will be trained using the same dataset. The batch size and optimizer utilized will differ. Finally, the outputs of both LeNet and AlexNet will predict four fundamental emotions. In addition, Google Colab is employed as the software platform [2].

II. VIDEO EMOTION DATABASE

An essential component of a video-based emotion detection system for training, evaluating, and testing the system is the database. There are five database characteristics. First, the subject's qualities. The database's attributes include the subject's gender, age group, race, facial hair, skin texture, face shape, and accessories. For example, the Japanese Female Facial Expression Database (JAFFE) contains solely Japanese women [8].

978-1-6654-8800-6/22/$31.00 ©2022 IEEE

Second, facial expressions, both spontaneous and nonspontaneous. Most face emotion databases have been compiled by acting out the emotion. The timing and temporal dynamics of these nonspontaneous facial expressions will vary from those of spontaneous facial expressions. Despite this, efforts for spontaneous facial expression are expanding since a great deal of research remains to be conducted. The MMI database is an example of a spontaneous facial expression [9].

The database's resolution is the third characteristic. The majority of the current facial expression has an excellent resolution. However, many applications of emotion identification need low-resolution photos or videos, such as video conferencing. This issue is solvable by subsampling the original dataset.

TABLE I. COMPARISON OF VARIOUS VIDEO EMOTION DATABASES

Database	Characteristics
CK+ [10]	• 593 sequences. • 123 subjects (18 to 50 years old) • Six emotions (happiness, sadness, anger, disgust, fear, and surprise). • The subject is American. • From 593 sequences, only 327 have emotion labels. • Image resolutions of 640×480 and 640×490
MMI [9]	• Web-based. • Six emotions (fear, happiness, anger, disgust, sadness, and surprise) • Static and dual view image • 69 subjects (19 to 62 years old) • Contains severe head pose variations • Image resolutions of 720×576 pixels.
ISED [7]	• 50 individuals of Indian (29 male and 21 female) • 428 video clips • Four emotions (happiness, surprise, sadness and disgust) • Resolution of 1920×1080.
JAFFE [8]	• 213 images of 7 emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral) • 10 subjects (Female Japanese) • 6 emotion adjectives by 60 Japanese subjects • Image resolution of 256×256
AR [11]	• 16 individuals (63 men and 53 women) • 13 conditions • 4 facial expressions (neutral, smile, anger, scream) • 3 illuminations (left light on, right light on, both lights on) • 6 occlusions (sunglasses, sunglasses/left light, sunglasses/right light, scarf, scarf/left light, scarf/right light) • Two sessions per person (2 different days). • Frontal view only. • Image resolution of 768×576 pixels
CAS-PEAL [12]	• 1040 individuals (595 men and 445 women) • Seven categories (accessory, lighting, background, pose, expression, distance) • Time/age consideration during image collection. • Consideration of surprise and open mouth categories. • Image resolution of 640×480 pixels
RaFD [13]	• 67 individuals (Caucasian males, Caucasian females, and Moroccan Dutch males) • Eight emotions (happiness, sadness, surprise, anger, disgust, fear, contempt, and neutral) • 3 gaze directions (left, straight, and right) • 5 camera angles (180°, 135°, 90°, 45°, and 0°) • Image resolution of 1024 × 681 pixels.

Fourth is the environment in which the photographs or videos were taken. Most available databases feature consistent, neutral backgrounds to facilitate emotion recognition. In addition, as in the real world, lighting conditions may not always remain constant, which poses a significant challenge to the emotion identification system. CAS-PEAL is a database of this type [12].

Face or head orientation comes fifth. Not many face expression databases are invariant to camera angle since it affects emotion recognition performance. The camera angle variation is seen in the MMI database [9].

Table I compares multiple video emotion databases. We chose the ISED database for our research because of its high image resolution, number of emotions, and number of individuals recorded.

III. VIDEO EMOTION RECOGNITION SYSTEM

The overall system flow is depicted in Figure 1. The system may be broken down into four distinct processes: initialization, feature extraction, emotion categorization, and final prediction.

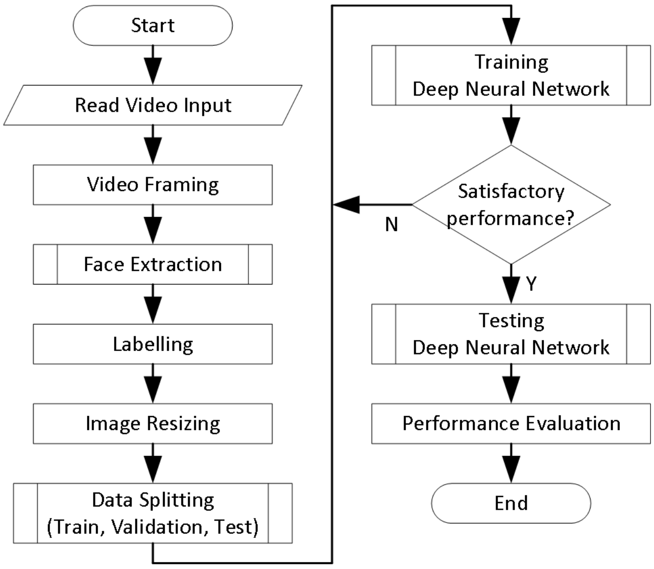

Fig. 1. Flowchart of the Video Emotion Recognition System

A. Preprocessing and Feature Extraction

Initialization is the beginning procedure, which consists of data entry till preprocessing. The input data will be evaluated to determine whether it is in video or picture format. As the data is in video format, it must frame the scene according to its frames per second before proceeding to the next phase, preprocessing. The library used is version 2.0 of OpenCV.

During preprocessing, the facial features of each frame are extracted and aligned using a face detector and a similarity transform. Haar Cascade is currently being implemented to extract faces from each video frame using the Viola-Jones method [14]. Positive and negative images are utilized in its training. Positive images are defined as those that the classifier must identify. While negative images are those that do not need to be detected by the classifier, positive images are those that do.

The Haar Cascade detects the face in each frame from input frames. The faces are cropped and saved to a different Google Drive folder using the OpenCV function. Although some faces in the frame may not be detected, they can be sorted once the process is complete.

B. Data Labelling and Splitting

The extracted face must then be labeled according to its emotion class, which includes disgust, happiness, sadness, and surprise. The labeling is performed by manually selecting the class emotion of the extracted faces based on their facial features, with the class folder serving as the destination. In addition, as shown in Table II, the labeled extracted face will be divided into three sections: training, validation, and testing.

TABLE II. DATA LABELLING AND SPLITTING

Emotion	Train	Validation	Test
Disgust	161	100	100
Happiness	519	344	120
Sadness	320	212	100
Surprise	111	73	50
Total	1111	729	370

C. Deep Learning Architectures

Transfer learning is an optimization that permits rapid advancement or enhanced performance while modeling the second task. Transfer learning is the improvement of learning in a new activity through transferring previously acquired knowledge from a related task. In this paper, two deep learning architectures, LeNet [15] and AlexNet [16], will be utilized for transfer learning because of their accuracy and popularity.

For LeNet and AlexNet, the input image is downsized to 28×28 and 227×227, respectively. In addition, image augmentation is applied to each dataset through the Keras function. The batch size will initially be set to 10. LeNet and AlexNet have total parameters of 61496 and 56377156, respectively. Because the dataset contains more than two classes, categorical cross entropy is used as the loss function for both architectures. While Stochastic Gradient Descent (SGD) is being implemented as an optimizer, performance metrics will be used to evaluate the architecture's accuracy.

Adjustments will be made to the parameters of both architectures in order to determine their effects. Optimizers and batch size will be modified as parameters. Adam and RMS prop will be substituted for the SGD optimizer. In the meantime, the batch size will increase from 10 to 100, making it ten times larger.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Experimental Setup

Google Colaboratory is used throughout the process as the integrated development environment. The cloud-based Jupyter Notebook environment is accessible via any web browser for free. Additionally, it permits the utilization of robust computing resources. Table III displays the Google Colab specifications utilized for this research.

TABLE III. GOOGLE COLAB SPECIFICATIONS

Hardware	Specification
GPU	Tesla T4
CPU	Intel Xeon CPU @ 2.20 GHz
RAM	13 GB
HD	37 GB

Indian Spontaneous Expression Database (ISED) is the dataset used, where the participant's race is Indian. It consists primarily of 29 males and 21 females. In addition, the dataset contains 428 video clips containing four emotions: disgust, happiness, sadness, and surprise.

The provided dataset is completely unlabeled and unclassified, indicating that it is in its raw form. Additionally, it is not divided into training, validation, or testing. Therefore, the dataset must be processed before implementation on the deep learning architecture. In addition, the ISED dataset is stored on Google Drive so that it can be imported into Google Colab during execution [2]. The storage must be unlimited because the output of the processed dataset will be enormous.

B. Feature Extraction Experiment

Figure 2 illustrates an example of the feature extraction procedure, including extracted video frames, the outcome of the Haar Cascade process, and an extracted face.

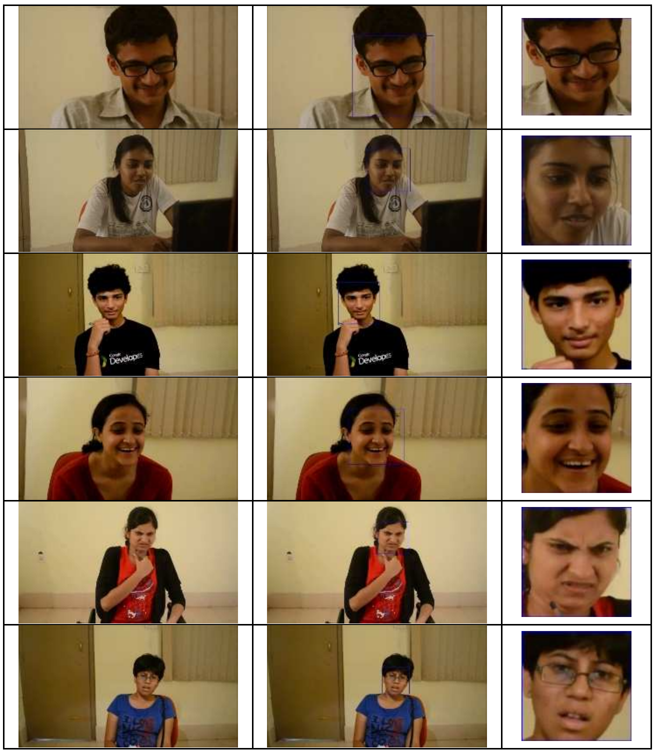

Fig. 2. Haar Cascade Feature Extraction

C. Transfer Learning Experiment

LeNet and AlexNet will contribute to this result, as both architectures use the same optimizer and batch size. The optimizer utilized is SGD, which was initialized to 0.001, and the batch size is ten. As shown in Table IV, AlexNet provided more accurate predictions than LeNet, which only achieved a 7.00 percent accuracy rate, which is very low. Additionally, the AlexNet loss is less than that of LeNet. One could say that the accuracy of a prediction is inversely proportional to the magnitude of the loss. As AlexNet is a deeper network than LeNet, the training processing time will be longer. Figure 3 displays the AlexNet confusion matrix.

TABLE IV. ACCURACY, LOSS, AND TRAINING TIME

Architecture	Accuracy (%)	Loss	Training Time (s)
LeNet	7.00	1.78	607
AlexNet	93.00	0.24	4625

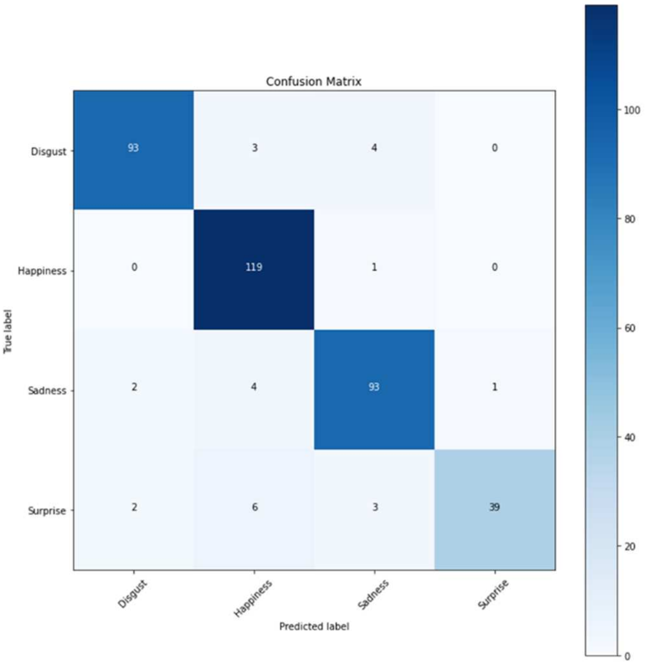

Fig. 3. Confusion Matrix of AlexNet

D. Optimizers Experiment

The current LeNet and AlexNet optimizer was initialized to SGD. To observe the effect, RMSprop and Adam are substituted. The objective is to investigate how optimizers contribute to optimal performance. Table V displays the outcomes of utilizing various optimizers. In contrast, Table VI demonstrates the accuracy of LeNet and AlexNet trained with different optimizers. Using the SGD optimizer, AlexNet achieved the highest accuracy level for the emotions of disgust and joy. As in the sadness class, it is tied with LeNet's Adam optimizer, which is also the best in the surprise class.

TABLE V. THE PERFORMANCE EVALUATION OF VARIOUS OPTIMIZERS

Arch	Metrics	SGD	RMSprop	Adam
LeNet	Accuracy (%)	7.00	85.00	84.00
	Loss	1.78	0.36	0.50
	Training (s)	607	1238	1351
AlexNet	Accuracy (%)	93.00	12.00	6.00
	Loss	0.24	9.68	2.38
	Accuracy	4625	5036	4982

TABLE VI. ACCURACY (%) OF LENET AND ALEXNET USING VARIOUS OPTIMIZERS

Arch	Optimizer	Disgust	Happiness	Sadness	Surprise
LeNet	SGD	7.00	95.00	68.00	6.00
	RMS	85.00	96.67	90.00	62.00
	Adam	84.00	95.00	93.00	90.00
AlexNet	SGD	93.00	99.17	93.00	78.00
	RMS	12.00	98.33	88.00	52.00
	Adam	6.00	95.83	67.00	8.00

In LeNet, RMSprop has a higher prediction accuracy than Adam, the difference being only 1%. The SGD optimizer is the lowest among these three optimizers by a significant margin. Nevertheless, its processing time is quicker than RMSprop and Adam, respectively. Consequently, LeNet will be implemented using the RMSprop optimizer. Despite this, AlexNet utilizing the SGD optimizer is still superior to the other two optimizers, as the difference in prediction accuracy is enormous. Additionally, it has the quickest training processing time. Therefore, AlexNet will continue to use the SGD optimizer.

E. Batch Size Experiment

It is evident from the previous section that LeNet using RMSprop as its optimizer yields the best results. Nevertheless, AlexNet with SGD remains the best architecture despite using alternative optimizers. This section will initialize the batch size to 100, which is ten times greater than the initial batch size. This is to determine the effect of batch size on the performance of deep learning architectures. Batch size is proportional to steps per epoch. Specifying the number of iterations per epoch that the deep learning architecture will execute is essential. To be clear, the size of the training dataset is 1111.

TABLE VII. THE PERFORMANCE EVALUATION OF TWO BATCH SIZES

Arch	Metrics	Batch Size	
		10	100
LeNet	Accuracy (%)	85.00	79.46
	Loss	0.36	0.53
	Training (s)	1238	1025
AlexNet	Accuracy (%)	93.00	71.08
	Loss	0.24	0.78
	Accuracy	4625	3767

As shown in Table VII, both LeNet and AlexNet experience a decline in prediction accuracy and processing time, while the losses increase when 100 batches are used. Thus, it could be argued that the larger the batch size, the fewer steps per epoch are required, resulting in a decrease in accuracy and a reduction in processing time.

TABLE VIII. ACCURACY (%) OF LENET AND ALEXNET USING DIFFERENT BATCH SIZES

Arch	Batch Size	Disgust	Happiness	Sadness	Surprise
LeNet	10	85.00	96.67	90.00	62.00
	100	47.00	97.50	84.00	30.00
AlexNet	10	93.00	99.17	93.00	78.00
	100	62.00	30.00	78.00	80.00

The accuracy of LeNet and AlexNet trained with different batch sizes is displayed in Table VIII. Implementing 100 batches increases the percentage accuracy of only the happiness class in LeNet. While the accuracy percentage of the other emotion classes decreased significantly. When 100 batches are applied, however, the accuracy of the surprise class trained with AlexNet increases marginally. As the accuracy percentage of the other emotion class decreased significantly.

TABLE IX. COMPARISON WITH THE BENCHMARK PAPER

Method	Accuracy
LeNet with RMSprop optimizer	85.00
AlexNet with SGD optimizer	93.00
LGBP Features using PCA+LDA classifier [7]	86.46

Table IX compares our proposed video emotion recognition system with the benchmark paper [7]. The benchmark paper utilized LGBP (Local Gabor Binary Pattern) and PCA+LDA (Principal Component Analysis + Linear Discriminant Analysis) classifiers for manual feature extraction. Consequently, our proposed AlexNet with SDG optimizer offers the most remarkable accuracy.

V. Conclusions and Future Works

We have presented the development of a transfer learning-based video-based emotion recognition system. The Indian Spontaneous Expression Database (ISED) was used for training, validation, and testing. Based on our experiments, AlexNet with the SGD optimizer and LeNet with the RMSprop optimizer achieved 93% and 85% recognition accuracy, respectively. The AlexNet architecture provides a higher recognition rate when compared to the benchmark paper. Future research may include the application of various deep learning architectures for transfer learning, using diverse databases and a variety of emotions.

Acknowledgment

The Ministry of Higher Education financially supported this research under the Fundamental Research Grant Scheme, FRGS19-076-0684 (FRGS/1/2018/ICT02/UIAM/02/4). The authors would also like to acknowledge the support from International Islamic University Malaysia and UIN Sunan Gunung Djati for the research facilities.

References

[1] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 443-449.

[2] T. S. Gunawan *et al.*, "Development of video-based emotion recognition using deep learning with Google Colab," *TELKOMNIKA (Telecommunication Computing Electronics and Control),* vol. 18, no. 5, pp. 2463-2471, 2020.

[3] Y. Fan, J. C. Lam, and V. O. Li, "Video-based emotion recognition using deeply-supervised neural networks," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 584-588.

[4] A. Latif, T. S. Gunawan, M. Kartiwi, F. Arifin, and H. Mansor, "Development of Image-Based Emotion Recognition using Convolutional Neural Networks," in *2021 IEEE 7th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, 2021: IEEE, pp. 47-52.

[5] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 445-450.

[6] A. Ashraf, T. S. Gunawan, F. D. A. Rahman, and M. Kartiwi, "A Summarization of Image and Video Databases for Emotion Recognition," in *Recent Trends in Mechatronics Towards Industry 4.0*: Springer, 2022, pp. 669-680.

[7] S. Happy, P. Patnaik, A. Routray, and R. Guha, "The Indian spontaneous expression database for emotion recognition," *IEEE Transactions on Affective Computing,* vol. 8, no. 1, pp. 131-142, 2015.

[8] M. Lyons, S. K. Akamatsu, M, and J. Gyoba, "Coding Facial Expressions with Gabor Wavelets," in *Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998: IEEE Computer Society, pp. 200-205.

[9] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE International Conference on Multimedia and Expo*, 2005: IEEE, p. 5.

[10] 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," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, 2010: IEEE, pp. 94-101.

[11] A. Martinez and R. Benavente, "The AR face database," CVC Technical Report 24, 1998.

[12] W. Gao *et al.*, "The CAS-PEAL large-scale Chinese face database and baseline evaluations," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans,* vol. 38, no. 1, pp. 149-161, 2007.

[13] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the Radboud Faces Database," *Cognition and Emotion,* vol. 24, no. 8, pp. 1377-1388, 2010.

[14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, 2001, vol. 1: IEEE, pp. I-I.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems,* vol. 25, 2012.