2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA) | 979-8-3503-2991-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAICTA59291.2023.10390247

Spontaneous Micro-expression Recognition using 3DCNN on Long Videos for Emotion Analysis

Budhi Irawan
School of Electrical Engineering
Telkom University
Bandung, West Java, Indonesia
budhiirawan@telkomuniversity.ac.id

Nugraha Priya Utama
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, West Java, Indonesia
utama@informatika.org

Rinaldi Munir
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, West Java, Indonesia
rinaldi@informatika.org

Ayu Purwarianti
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, West Java, Indonesia
ayu@informatika.org

***Abstract*— Research related to the recognition of micro-expression continues to develop yearly. New ways are always found, from the preprocessing stage and feature extraction to the classification stage, with learning models to recognize emotions and measure their performance. Based on previous studies, many problems still need to be encountered in recognizing micro-expression in long video sequences. This study proposes a new framework for the introduction of micro-expression, which consists of several stages of the process, namely the stages of preprocessing tiered novels and the Three Dimensional Convolutional Neural Network (3DCNN) learning model and ends with measuring its accuracy. By implementing this preprocessing stage, the micro-expression features found on the face are protected from damage and loss. In addition, some experiments on the 3DCNN learning model have increased the accuracy of micro-expression recognition. We used the SMIC dataset to evaluate the capabilities of the proposed micro-expression recognition framework. The accuracy results obtained in this dataset are 81.70%. These results demonstrate the effectiveness of the proposed new micro-expression recognition framework.**

Keywords—Micro-expression; 3DCNN; deep learning

I. INTRODUCTION

Micro-expressions are facial expressions that last very briefly, are involuntary, and occur in some regions of the face. Compared to long and noticeable changes from regular facial expressions, micro-expression has a short period of not more than 1 second. Subtle changes in intensity due to facial brawn gestures caused by micro-expression appear only to a small extent and occur in only a few areas of the face [1], [2]. Spontaneous micro-expression can disclose actual emotions and expose someone's lies. Based on this fact, it is essential to recognize micro-expression for various purposes [3], [4], [5], such as for state security systems, detecting lies, interrogating criminal cases by the police, diagnosing diseases by doctors, business negotiations, psychoanalysis, and others. As a result of the concise duration of occurrence and the subtle changes in the muscles in the facial area, it is difficult for a person to discover and recognize micro-expressions. Nay, an expert trained by an expert micro-expression training kit [6] needs help carrying out the micro-expression recognition process.

Several approaches have been made to recognize micro-expression by modelling the delicate alteration of micro-expression in the spatial region [7]. There are two main approaches to the micro-expression recognition process. The first stage is preprocessing and facial feature extraction from video clips in the dataset. Next, the second stage recognizes micro-expressions around the facial area using a classifier to classify the features extracted in the previous stage. The activity of recognizing micro-expressions is a task of recognizing specific patterns among well-known classifier methods such as Support Vector Machine (SVM) [8], [9], [10], which are widely used and implemented to recognize micro-expressions.

Apart from micro-expression, there are also other expressions on the human face, namely macro-expression. Macro-expression has a duration that is more prolonged and more intense. Recent advances in deep learning have shown that its popularity is very significant, especially in recognizing micro-expressions and detecting micro-expressions in finding the peak frame of a long video sequence has yet to be well resolved [11]. Thus, there is a challenge in tracking down micro-expressions in long video sequences. Facial micro-expression generally experiences three phases: the onset, apex, and offset. The onset phase is the first phase when the facial brawn starts to contract. Then, the apex phase is the phase where the facial motion is at its zenith intensity, and the offset phase indicates the phase when the facial brawn returns to an impartial state.

II. RELATED WORK

CNN has shown considerable superiority in various fields, outperforming handcrafted methods and traditional classifiers. Research [4] proposes an approach to recognize micro-expression with deep recurrent convolutional networks to obtain the spatiotemporal flow of micro-expression frame sequences. The experiment in this study is to apply the CNN model, which consists of a set of recurrent convolutional layers to perform feature extraction and classification processes in recognizing micro-expressions. The paper [12] proposes a method to study features using the latest distance network. This proposal is carried out by recognizing micro-expression with multiple views. This method first extracts the scale-invariant feature corresponding to a set of landmark area points on each face

frame. Another study presented a deep fusion convolutional neural network to introduce two-dimensional and three-dimensional micro-expression recognition. The architecture includes feature extraction stages, feature fusion and softmax layers [13].

This article [14] proposes to utilize three-dimensional stream-based CNN to analyze the micro-expression of video clips. All features are entirely extracted from some selected datasets from every smooth muscle movement on the face. It takes many datasets to implement this deep learning method. If the number of datasets is tiny, it will impact obtaining limited learning outcomes, and ultimately, the resulting level of accuracy is not as expected. A dual-stream temporal-domain design was proposed to improve micro-expression recognition accuracy [15]. This research focuses on how the feature extraction process becomes essential in helping the classification process produce better accuracy.

Methods for extracting conventional features are still widely used to recognize patterns in images or objects. There are several drawbacks to the feature extraction process using conventional methods in micro-expressions recognition research, especially in areas that are difficult to extract in detail. In research, [16] proposed a method that can effectively recognize micro-expression using conventional feature extraction methods. In addition, recognizing expressions in short video clips and low light intensity levels is still challenging in recent research. Various ways to overcome this problem continue to be explored and developed. One of them is done in the study [17], which proposes a method to study the problem of adaptation level in recognizing micro-expression. Several deep learning models have been specialized to address micro-expression recognition problems caused by the limited number of video-based training examples. As a result of the limited sample and unbalanced emotion classes, CNN only works optimally because deep learning usually needs to learn many parameters. Until now, new methods are continuously being developed to get better results[18], especially to answer the challenge of detecting frames that contain micro-expressions from the onset, the apex and the offset frame.

Various suggestions and the development of new methods for recognising micro-expression continue to emerge for better accuracy. One of them is the proposed CNN learning model [19] to carry out the face detection process, where the Eulerian Video Magnification method is used to enhance the capability to recognise micro-expressions. The CNN classifier is used to classify micro-expressions previously detected into one of seven universal emotion classes. Other studies [20] and [21] proposed using three-dimensional CNNs to increase accuracy in recognising micro-expressions in various proposed frameworks. One is the proposed architecture to recognise micro-expressions using three-dimensional neural networks (3DCNN). The architecture consists of a 3D convolution layer, pooling layer, batch normalisation, and other added layers to construct the model architecture. Another architecture is to propose a combination of two-dimensional and three-dimensional CNN networks to be able to recognise micro-expressions. The architecture consists of a network that can extract spatiotemporal features built with one-dimensional and two-dimensional convolution layers.

Improvement of the 3DCNN learning model continues to be carried out in various ways to enhance micro-expression recognition accuracy. This paper [22] proposes the architecture of the CNN learning model by utilizing an optical flow-based feature extraction method to obtain complete and valuable features for micro-expression spotting. In this research, a process was also carried out to increase the limited number of datasets using data augmentation and pseudo-labelling methods. Further studies [23] propose a learning model architecture that simultaneously performs feature extraction processes between spatial and temporal domains. The model used still applies three-dimensional CNN. This study focuses on determining the apex frame as a reference in the micro-expression recognition process. In this way, it is constructive to increase the level of recognition accuracy. The most recent paper [18] presents experimental results from applying a three-dimensional CNN architectural model to high-resolution micro-expression video clip datasets. It carries out the identification process of adding dropout layers that vary in value and number, tested on the existing learning model architecture and researched before.

III. METHODOLOGY

In this research, we develop a new framework that can recognize micro-expressions. The framework was developed to deal with the limitations of spontaneous micro-expression datasets, recognize micro-expressions and analyze emotions from micro-expression that occur with a better performance accuracy than previous research. Figure 1 shows the design of the proposed new framework developed from this research.

The stages of developing a new framework that will be completed in this research are a novel tiered preprocessing stage, which begins with collecting and selecting spontaneous micro-expression datasets in the form of raw video. Then, proceed with grouping the dataset based on the emotion class. This preprocessing stage includes image processing by following the following steps: video conversion into a series of sequential frames, face detection, face alignment, face landmark detection, face cropping, image resizing, eye masking, RoI Combining, and grey scaling. The next step is the classification stage using the deep learning architecture, namely MicroExpSTCNN [24], which has been modified and parameter setting. The 3DCNN method is a method that has a CNN basis. 3DCNN has proven effective in capturing motion information from continuous video frames and can capture any information stored in each feature that flows spatiotemporal.

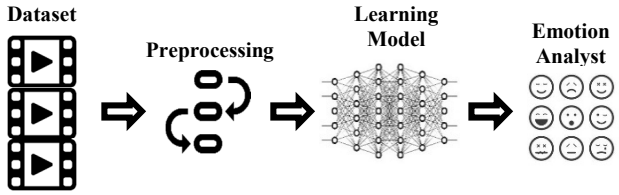

Fig. 1. Proposed new framework

A. Dataset

There are two micro-expression data sets based on manufacturing scenarios: spontaneous and posed. The main difference between spontaneous micro-expression and posed types lies in the pertinence of facial movements and actual

emotional states. In micro-expression of the posed type, facial expressions are intentionally displayed but are irrelevant to the current emotional state. Therefore, micro-expression of this pose could be more helpful for recognizing actual emotions. Meanwhile, spontaneous micro-expressions are facial expressions that are the same as the actual or underlying emotional state.

This study used one spontaneous dataset, namely the SMIC dataset of the High Speed (HS) type [25]. The two annotators labelled this dataset according to each participant's self-reported emotions. The annotator follows the FACS guide by checking every frame of video previously recorded. In this dataset, a video database was built that describes a person's spontaneous emotional state and can be used for research in micro-expression recognition to analyze the emotions that occur. Some video clips are provided for participants to watch to trigger strong emotions. The participants were asked not to show emotions that arise from actual feelings on their faces. The participants were asked to commit to the previously agreed rules. Figure 2 shows a sample frame of a participant on the SMIC dataset.

Fig. 2. Sample of some frames on the SMIC dataset

Each candidate apex frame found will be compared with the emotions delivered by each participant. Only labels unchanging with each participant's report will be marked in the dataset. This dataset consists of 164 video clip files for 9-343 seconds. There were 20 participating subjects aged 22-34 years, consisting of 6 women and 14 men. Participants consisted of various ethnic groups, including 10 Asians, 9 Europeans and 1 African. Ten participants wore glasses with a resolution of 640x480 pixels and a video frame rate of 100fps.

Table I shows the emotion class of the dataset and the number of video clip samples. This study uses a dataset ratio of 80% for training and 20% for validation data. Note that the separation of training and validation data is executed only once. Then, the same training and validation data were used for each observation across all experiments.

TABLE I. SUMMARY OF THE RATIO OF THE TRAINING SET AND VALIDATION SET FROM THE SMIC DATASET USED

Emotion Classes	Total Number	Ratio	
		Training Set	Validation Set
Negative	70	56	14
Positive	51	41	10
Surprise	43	34	9
Total	164	131	33

B. *Preprocessing*

To recognize spontaneous micro-expressions and analyze emotions from a long of videos on faces, a tiered preprocessing stage is proposed, contributing to this study. Preprocessing begins with classifying the emotional classes from the dataset into three emotional classes: negative, positive, and surprise classes. The first preprocessing stage is converting video clips from the dataset into a series of sequential frames with a resolution of 640x480 pixels, as shown in Figure 3. Next, the face detection process is carried out on the frame and face alignment, which is a technique in which the face image frame is rotated according to the corner of the eye so that the left eye and the right become aligned on the horizontal axis. Then, detect the AU point on the face using 68 facial landmarks. In the next preprocessing stage, only the face is cut, and the left and right eye areas are covered with a polygon-shaped mask [26]. Then, cut the left and right eyes and mouth according to predetermined landmark boundaries, recombine them to form a dimension of 128x128 pixels, and end with the greyscaling process.

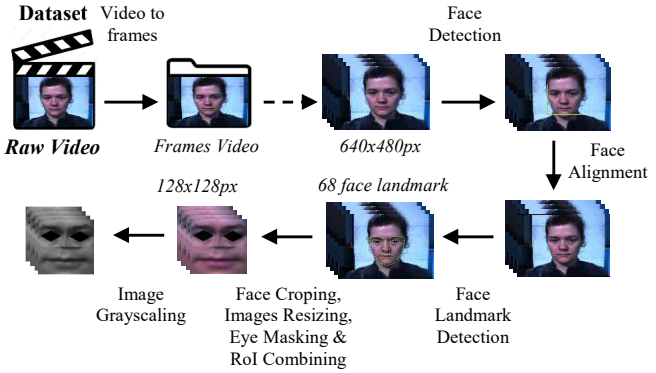

Fig. 3. Proposed new tiered preprocessing stages

C. *MicroExpSTCNN Experiments and Modifications*

After the preprocessing stage, the next stage is the training and classification stage by applying the CNN learning model architecture. CNN is a type of neural network widely used in image processing. This study uses the MicroExpSTCNN model architecture, which has been modified previously. The following is the built architecture shown in Figure 4.

The architecture of the MicroExpSTCNN model is proposed by utilizing spatiotemporal features during micro-expression in video frame sequences. The input dimensions for this proposed model are w × h × d or 128 x 128 x 10, where w and h are the size of the video frame or spatial data, 128x128 pixels. The value of d is the depth or time of the sequence of some video frames, which is the apex occurrence interval depending on the datasets used. The proposed model architecture consists of a 3D convolution layer, 3D max pooling layer, flatten layer, dropout layer, softmax layer, dense layer, activation function and fully connected layer.

The proposed model results from modifying the MicroExpSTCNN model [24]. The main difference is the addition of the 3Dmax Pooling layer to become two layers. The purpose of adding this layer is to create a robust feature extraction method by taking the maximum value from each 3D sub-block pooling to identify the most significant features in the

data volume, which can help the model to focus on relevant information. In addition, three dropout layers were added to help increase the model's generalization. This means the model can make better predictions on data that is not visible during training, thereby reducing the risk of overfitting.

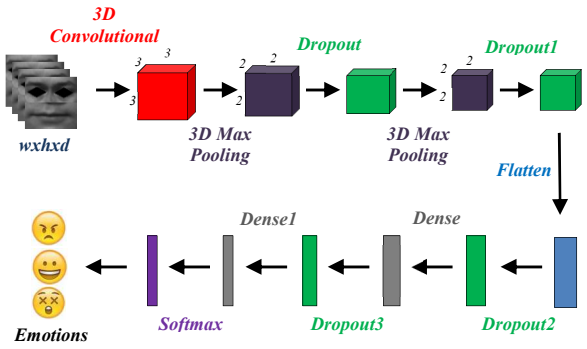

Fig. 4. MicroExpSTCNN Architectures Modifications

The 3D convolution layer extracts spatiotemporal features by applying the convolution stage using the 3D kernel. 3DCNN can use filters in a spatiotemporal way. The 3D max pooling layers progressively decrease the dimensional output of the 3D convolution layers while retaining essential features. The 3D max pooling layer selects the best feature representation in the spatiotemporal window. The use of dropouts in this network is to reduce the overfitting of the model. Dropouts are used to add regularised capabilities to the proposed network. The flatten layer is a layer that has the function of changing input with many dimensions into a one-dimensional array required for a fully connected layer. The fully connected layers are required to recognize more non-linearity in a network architecture. The softmax layer generates class values from emotions according to the number of classes.

The proposed model architecture consists of one 3D convolution layer with 32 filters with 3x3x3 dimensions, two 3D max pooling layers which have kernel sizes of 3x3x3 and 2x2x2 respectively, four dropout layers, one flatten layer, two dense layers and one softmax layer. The softmax layer's final dimensions depend on the dataset's number of expression labels. Table II summarises the proposed model architecture regarding the filter and output dimensions of the sundry layers used.

TABLE II. SUMMARIZES THE PROPOSED MODEL ARCHITECTURE IN TERMS OF FILTER AND OUTPUT DIMENSIONS OF THE VARIOUS LAYERS

Layer Type	Filter	Filter Size	Output Dimension
Input	-	-	128x128x10
3D Convolution	32	3x3x3	32x126x126x8
3D Maxpooling	-	3x3x3	32x42x42x2
Dropout	-	-	32x42x42x2
3D Maxpooling	-	2x2x2	32x21x21x1
Dropout 1	-	-	32x21x21x1
Flatten	-	-	14112
Dropout 2	-	-	14112
Dense	-	-	32
Dropout 3	-	-	32
Dense 1	-	-	3

D. HyperParameter Setting

After the preprocessing stages are carried out, the dataset is stored in a pickle format and entered into the learning model with several parameters observed and configured first, including determining depth. The depth here means that the number of micro-expression video frames containing a more significant number of apex frames is the same as the number of apex frames corresponding to the data in the annotation file provided by the dataset provider to maintain data consistency. Then, divide the dataset into training and validation data with a ratio of 80:20.

This study defines the model in the Keras library with Tensorflow as the backend. Models are trained and tested on Google Colab Pro. This model applies the Categorical Cross Entropy loss function. The optimizer used is ADAM, which is different from the baseline, which uses Stochastic Gradient Descent (SGD). Adam is an optimization algorithm that can substitute for the classical SGD operation to renew weights based on trained data. [27] iteratively. The learning rate values set in this experiment are {0.01, 0.001, and 0.0001}, dropout = {0.3, 0.5, 0.7}, and epoch = {80, 100, 120} with batch size = {20}. As an initial setting, the system is run using learning rate = 0.01, dropout = 0.3, batch size = 20, and 100 epochs.

A typical evaluation metric for recognizing micro-expressions is measuring accuracy. Accuracy metrics generally measure the ratio of correct predictions to the total samples evaluated. Mathematically, it is presented in Equation 1.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (1)$$

IV. RESULT AND DISCUSSION

In the early stages of the experiment, several parameters were observed, including learning rate, dropout, and epoch observations. Learning rate functions control how fast the model learns the problem. A learning rate that is too small can slow the learning process, whereas if it is too large, it can cause the model to converge too quickly to a suboptimal solution. From Table III, a learning rate of 0.001 gives the best performance. Increasing the learning rate to 0.01 causes learning to be too fast, while decreasing the learning rate to 0.0001 causes learning to be too slow.

TABLE III. ACCURACY OF LEARNING RATE OBSERVATION RESULTS

Learning Rate	Accuracy (%)
0.01	76.81
0.001	81.70
0. 0001	73.48

During training, overfitting often occurs due to the small number of datasets and unbalanced class composition. So, a dropout layer is added to prevent the model from learning too strong a dependency on a particular training dataset, which could result in overfitting. Dropout is one of the techniques used to reduce overfitting in deep learning models, including the 3DCNN model, and dropout is a regularization method that randomly turns off several neurons in hidden layers during the training process. In this experiment, dropout does not improve

model performance but can help overcome overfitting. Table IV shows that the best performance is obtained at the smallest dropout value.

TABLE IV. ACCURACY OF DROPOUT OBSERVATION RESULTS

Dropout	Accuracy (%)
0.3	81.70
0.5	76.17
0.7	68.90

According to the baseline method, the model was trained for 100 epochs. In this experiment, the performance was quite good. After the epoch is added, the performance decreases, and overfitting occurs. Besides that, it is also done by giving an epoch value of 80, which results in decreased performance and overfitting. The best model is obtained with a learning rate parameter value of 0.001, a dropout of 0.3, and 100 epochs, as shown in Table V.

TABLE V. ACCURACY OF EPOCH OBSERVATION RESULTS

Epoch	Accuracy (%)
80	74.21
100	81.70
120	78.37

To verify the effectiveness and superiority of this research, a comparison was made with previous studies. Table VI shows that this study produced a preferable level of accuracy than the results of previous studies, namely 81.70%. This means that a novel tiered preprocessing stage and the proposed learning architectural model can work according to the plan to increase micro-expression recognition accuracy in long videos. A novel tiered preprocessing stage produces input frames with essential features and minimizes the loss of features needed in the learning stage. The eye masking process and cutting three facial areas, namely the left eye and eyebrow, the right eye and eyebrow, and the mouth and combining them into a frame with a size of 128x128 pixels helps to reduce interference with the classification process so that only the essential features recognition process is carried out.

TABLE VI. MICRO-EXPRESSION RECOGNITION ACCURACY COMPARISON ON SMIC DATASETS

Method	Accuracy (%)
3DFlow CNN [14]	55.49
STRCNN [4]	72.30
MESTCNN [28]	68.75
DSTICNN64 [15]	78.78
MER3DCNN [20]	76.92
MSAF [16]	73.60
AMAN [17]	79.87
ODCNN [29]	74.80
MicroExpSTCNN [24]	68.75
3D-CNNMED [18]	80.94
Ours	**81.70**

In addition, the proposed MicroExpSTCNN learning architecture model has been modified and added to several layers, including a 3DConvolution layer with a filter size of 3x3x3 and the addition of two layers of 3DMaxPooling with a filter size of 3x3x3 and 2x2x2 is enough to help improve accuracy when introducing micro-expression spatiotemporal. In contrast, the addition of one dropout layer serves to reduce overfitting. The performance of the built model also depends on network hyper-parameter settings. One setting that is quite important is the filter dimensions used for 3DConvolution. The impact of using 3D kernel size is that it can increase spatiotemporal exploitation in the feature extraction process. At this stage, experiments have been conducted to find the maximum filter size to achieve the best results.

V. CONCLUSION

This study developed a new framework with several processing stages to recognize micro-expressions and analyze the occurring emotions. The test results show that the proposed method is excellent compared to the baseline method, and several previous studies have resulted in an accuracy of 81.70% in the SMIC dataset. That shows the successful use of a novel tiered preprocessing stage and the 3DCNN learning model. The new framework that has been implemented has provided changes and increased micro-expression recognition accuracy. Although it was still encountered during the initial experiment, the model needed to be more balanced due to the limited number of datasets.

ACKNOWLEDGMENT

The first author is a Telkom Foundation of Education employee as a lecturer at the School of Electrical Engineering, Telkom University. Now, he is pursuing a doctoral program at the School of Electrical Engineering and Informatics, Bandung Institute of Technology. Telkom University supports this work.

REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, pp. 39–58, 2009.

[2] X. B. Shen, Q. Wu, and X. L. Fu, "Effects of the duration of expressions on the recognition of microexpressions," *J. Zhejiang Univ. Sci. B*, vol. 13, pp. 221–230, 2012.

[3] Y. Guo, Y. Tian, X. Gao, and X. Zhang, "Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method," *Int. Jt. Conf. Neural Networks*, pp. 3473–3479, 2014.

[4] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal Recurrent Convolutional Networks for Recognizing Spontaneous Micro-expression," *IEEE Trans. Multimed.*, vol. 22, pp. 626–640, 2020.

[5] M. A. Takalkar and M. Xu, "Image based Facial Micro-Expression Recognition using Deep Learning on Small Datasets," in *International Conference on Digital Image Computing: Techniques and Applications*, 2017, pp. 1–7.

[6] J. Endres and A. Laidlaw, "Micro-expression recognition training in medical students: A pilot study," *BMC Med. Educ.*, vol. 9, pp. 1–6, 2009.

[7] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, and G. Zhao, "Towards Reading Hidden Emotions : A Comparative Study of Spontaneous Micro-expression Spotting and Recognition Methods," *IEEE Trans. Affect. Comput.*, vol. 30, pp. 1–14, 2017.

[8] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 1449–1456. doi: 10.1109/ICCV.2011.6126401.

[9] Y. Liu, J. Zhang, W. Yan, and S. Wang, "A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition," *IEEE Trans. Affect. Comput.*, pp. 1–12, 2015, doi: 10.1109/TAFFC.2015.2485205.

[10] S. T. Liong, J. See, K. S. Wong, and R. C. W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Process. Image Commun.*, vol. 62, pp. 82–92, 2018, doi: 10.1016/j.image.2017.11.006.

[11] Y. H. Oh, J. See, A. C. Le Ngo, R. C. W. Phan, and V. M. Baskaran, "A survey of automatic facial micro-expression analysis: Databases, methods, and challenges," *Front. Psychol.*, vol. 9, pp. 1–45, 2018, doi: 10.3389/fpsyg.2018.01128.

[12] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A Deep Neural Network-Driven Feature Learning Method for Multi-view Facial Expression Recognition," *IEEE Trans. Multimed.*, vol. 18, pp. 2528–2536, 2016.

[13] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2D+3D Facial Expression Recognition with Deep Fusion Convolutional Neural Network," *IEEE Trans. Multimed.*, vol. 19, pp. 2816–2831, 2017.

[14] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3D flow convolutional neural network," *Pattern Anal. Appl.*, vol. 22, no. 4, pp. 1331–1339, 2019, doi: 10.1007/s10044-018-0757-5.

[15] W. Zhu and Y. Chen, "Micro-expression recognition convolutional network based on dual-stream temporal-domain information interaction," *Proc. - 2020 13th Int. Symp. Comput. Intell. Des. Isc. 2020*, pp. 396–400, 2020, doi: 10.1109/ISCID51228.2020.00096.

[16] M. Wang, "Micro-expression Recognition Based on Multi-Scale Attention Fusion," *Proc. 2021 IEEE Int. Conf. Data Sci. Comput. Appl. ICDSCA 2021*, pp. 853–861, 2021, doi: 10.1109/ICDSCA53499.2021.9650164.

[17] M. Wei, W. Zheng, Y. Zong, X. Jiang, C. Lu, and J. Liu, "a Novel Micro-Expression Recognition Approach Using Attention-Based Magnification-Adaptive Networks," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2022-May, pp. 2420–2424, 2022, doi: 10.1109/ICASSP43922.2022.9747232.

[18] W. S. P. Bayu and A. Setyanto, "3D CNN for Micro Expression Detection," *ICOIACT 2022 - 5th Int. Conf. Inf. Commun. Technol. A New W. to Make AI Useful Everyone New Norm. Era, Proceeding*, pp. 397–401, 2022, doi: 10.1109/ICOIACT55506.2022.9972194.

[19] S. C. Ayyalasomayajula, B. Ionescu, and D. Ionescu, "A CNN Approach to Micro-expression Detection," in *SACI 2021 - IEEE 15th International Symposium on Applied Computational Intelligence and Informatics, Proceedings*, 2021, pp. 345–350.

[20] Y. Jiao, M. Jing, Y. Hu, and K. Sun, "Research on a micro-expression recognition algorithm based on 3D-CNN," in *2021 3rd International Conference on Intelligent Control, Measurement and Signal Processing and Intelligent Oil Field, ICMSP 2021*, IEEE, 2021, pp. 221–225.

[21] L. Wang, J. Jia, and N. Mao, "Micro-Expression Recognition Based on 2D-3D CNN Lin," in *Proceedings of the 39th Chinese Control Conference July 27-29, 2020, Shenyang, China Micro-Expression*, 2020, pp. 3152–3157.

[22] C. H. Yap *et al.*, "3D-CNN for Facial Micro- and Macro-expression Spotting on Long Video Sequences using Temporal Oriented Reference Frame," *Comput. Vis. Pattern Recognit.*, pp. 7016–7020, 2022, doi: 10.1145/3503161.3551570.

[23] K. K. Talluri, M. A. Fiedler, and A. Al-Hamadi, "Deep 3D Convolutional Neural Network for Facial Micro-Expression Analysis from Video Images," *Appl. Sci.*, vol. 12, no. 21, 2022, doi: 10.3390/app122111078.

[24] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks," *Int. Jt. Conf. Neural Networks*, no. July, pp. 1–8, 2020.

[25] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A Spontaneous Micro-expression Database: Inducement, collection and baseline," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*, 2013, pp. 1–6. doi: 10.1109/FG.2013.6553717.

[26] G. B. Liong, J. See, and L. K. Wong, "Shallow Optical Flow Three-Stream Cnn for Macro- and Micro-Expression Spotting From Long Videos," *Proc. - Int. Conf. Image Process. ICIP*, vol. 2021-Septe, pp. 2643–2647, 2021, doi: 10.1109/ICIP42928.2021.9506349.

[27] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015, pp. 1–15.

[28] S. Prasanna, T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous Facial Micro-Expression Recognition using 3D Spatiotemporal Convolutional Neural Networks," in *IJCNN 2019. International Joint Conference on Neural Networks.*, IEEE, 2019, pp. 1–8.

[29] K. S. Min, M. Asyraf Zulkifley, and N. A. Mohamed Kamari, "Optimized Dense Convolutional Neural Networks for Micro-expression Recognition," *2022 12th IEEE Symp. Comput. Appl. Ind. Electron. ISCAIE 2022*, pp. 288–293, 2022, doi: 10.1109/ISCAIE54458.2022.9794470.