2023 The 23rd International Conference on Control, Automation and Systems (ICCAS 2023)
Yeosu Venezia Hotel & Utop Marina Hotel, Yeosu, Korea, Oct. 17∼20, 2023

A Slim Convolutional Neural Network for Low-cost Masked Facial Expression Classifier

Muhamad Dwisnanto Putro[1], Vecky Canisius Pokoel[2], Adri Priadana[3], Jinsu An[4] and Kang-Hyun Jo[5*]

[12]Department of Electrical Engineering, Sam Ratulangi University, Indonesia (dwisnantoputro@unsrat.ac.id, vecky.poekoel@unsrat.ac.id)

[345*]Department of Electrical, Electronic, and Computer Engineering, University of Ulsan, Korea (priadana3202@mail.ulsan.ac.kr, jinsu5023@gmail.com, acejo@ulsan.ac.kr) * Corresponding author

Abstract: The facial expression emphasizes a predictive facial gesture widely used for emotion recognition. The pandemic situation encourages the demand for the use of masks in specific areas, urging a challenge from a facial analysis method to accurately predict limited facial features. Practical applications demand a classification system that operates quickly. This paper proposes an accurate real-time facial expression recognition for masked faces. It offers an efficient backbone that improves from MobileNetV2, which drives the model to generate parameters and less computation. The represented context module is introduced to capture the attention of interest information and highlight features globally without placing a significant computational burden. In order to emphasize a reliable classification system in real applications, it proposes the development of a masked facial expression dataset, namely M-KDEF. The experimental results show that the facial expression classifier performs compared to other methods with an accuracy of 96.99% on the M-KDEF and 87.14% on the M-LFW dataset. It also examines the speed of the integrated system that obtains a data processing speed of 61.30 frames per second on the CPU.

Keywords: Facial expression, facial masked, efficient model, CPU, real-time.

1. INTRODUCTION

Facial expression recognition (FER) is a cognitive method to identify the facial gesture representing human feeling. This method tends to recognize the distinctive facial features that describe facial emotions [1]. The uniqueness of each of these expressions facilitates an approach to paying more attention to facial regions for predictive decisions. The facial analysis methods' challenge is encouraging a vision system to predict accurately. One of them is the masked occlusion used, which provides limited information on facial features.

In a pandemic situation, industrial and medical places require faces to wear masks [2]. Therefore, the need for a masked facial expression recognition system becomes a crucial issue by testing a classification system to work accurately, which minimizes prediction errors. This challenge offers a partially covered face view. In general, masks ignore essential information on the mouth, cheeks, and nose. However, these features provide crucial information that supports the predictive stage for facial expression work. Expressions of happiness, surprise, fear, and disgust display distinctive gestures on these facial elements [3]. Therefore, some studies are constrained by achieving low accuracy due to the high rate of false positives [4], [5], [6].

The Convolutional Neural Network method has shown robust results for extracting important features [7]. This approach utilizes weighted kernels to filter information from the spatial area for each input feature map. The performance gradient updates the kernel weight and drives the network to produce a minimal loss score. The performance of this method makes it possible to get feature extraction with proper quality even though some objects are occluded [8]. Therefore, the work of this paper applies CNN as a feature extraction that accurately discriminates limited features to improve predictive performance. In addition, an attention module can help a deep learning network to obtain global relationships from a set of spatial regions [9]. It can improve network performance by capturing essential features and reducing trivial features. The global-based Attention module can capture a wide range of information and find feature relationships by generating similarity maps from all spatial positions.

The superiority of CNN in extracting information is not balanced with its efficiency factor. It weakens a deep learning method to be implemented on low-cost devices. Even robotic devices use CPUs to process their supporting instruments [10]. It illustrates that practical applications require a vision system to operate smoothly in real-time. Therefore, the work of this paper builds a vision system to recognize masked facial expressions using an efficient architecture. It applies a shallow convolution layer while maintaining its accuracy. To improve prediction accuracy, it proposes a represented global context module that can capture specific features globally.

Based on the previously mentioned problems, the major contributions of this paper are as follows:

1. A slim backbone architecture is improved from MobileNetV2 that efficiently extracts and discriminates the limited facial feature using an inverted residual block.
2. A represented global context module is presented to highlight the specific facial element and pays great attention to essential features that increase prediction accuracy.
3. The recognition system achieves high performance and to other methods. It can perform at a real-time processing speed of 204 FPS on a CPU-based device.

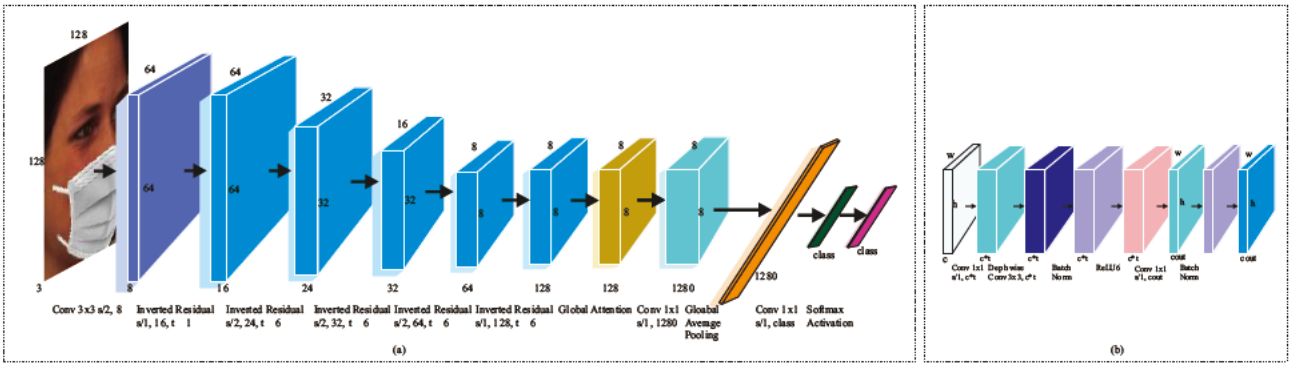

Fig. 1.: (a) The proposed architecture. A backbone module is employed to screen facial features using (b) a slim inverted residual module. It uses a superficial layer to boost the speed in the inference phase.

2. PROPOSED ARCHITECTURE

The proposed architecture contains three main modules, as shown in Fig. 1. First, a slim backbone comprehensively extracts critical facial features with an inverted residual block. Secondly, the represented global context captures the contextual information related to the expression. Lastly, the classifier module predicts the facial expression categories.

2.1 Backbone Module

The extractor feature is an essential process to catch out the distinctive information. It can comprehensively extract the input features map using specific operations. A convolutional Neural Network is a robust extractor module that delivers a satisfying result to filter information and generates accurate prediction [11]. The proposed backbone is used a CNN approach considering the efficiency of operating in real-time speed in inexpensive devices. Based on this motivation, the proposed network adopts an inverted residual module [12], which uses light operation. Fig 1 (b) presents that this module uses two simple 2D-convolutional and a depth-wise layer. It extracts the input features $x_i \in R^{h\times w\times c}$ on the sequential operation until it reaches the output block. The module followed a Relu activation and two Batch Normalization, as defined

$$y_i = C_{1+BN}(\delta(C_{DW+BN}(C_1(x_i)))), \qquad (1)$$

where C_1 is a convolutional operation using 1×1 weighted filter to extract channel-based information while expanding the variety of features along the channel size. C_{DW+BN} is Depth-wise convolutional using 3×3 weighted filter with Batch Normalization that efficiently extracts features in the spatial dimension. This operation is lighter than standard convolution, so that it can produce fewer parameters. C_{1+BN} is applied in the last module to extract the channel-based features using 1×1 kernel with Batch Normalization. It applies δ as ReLU activation to prevent the overfitting issue.

The proposed network uses an image with 128×128 in an RGB channel. It applies 2D-convolutional with 3×3 kernel in the beginning phase to reduce the dimension as well as extract the features. This process is more robust than the pooling operation. Furthermore, the proposed network employs six times the inverted residual with a different number of channels. It applies six scale expansion channels in most phases except in the first stage, which only uses one factor.

2.2 Represented Global Context Module

In general, the attention module has contributed to improving performance by capturing specific features of objects. This module also strengthens relationships between objects by finding similarities in the location of spatial regions. The proposed attention module consists of global contextual (GC_i)and reconstruction blocks (RC_i). The global contextual block captures the feature of interest from a feature map ($x_i \in R^{h\times w\times c}$) obtained from the channel representation and the relation between long-range information. This module can be formulated as follows:

$$GC_i = C_3(GAP(C_7(x_i) \otimes C_3(x_i) \otimes C_3(x_i))) + x_i \quad (2)$$

where C_7 and C_3 is 2D-convolutinal using 7×7 and 3×3 filter, respectively. It uses a big kernel to generate a single feature map that summarizes the important facial information along the channel. It also produces half feature maps to efficiently the computation in the multiplication process. The vectored global feature is generated from the global average pooling GAP of multiplication between single and half feature maps, constructing $c/2$ dimension. Furthermore, represented global information is multiplied using dot product operation to half feature maps. It also applies standard convolution in the last phase to restore the feature variation from half-compressed and returns the same channel size as the input feature.

The reconstruction module is employed after the output of the attention module. This block re-extracts features efficiently by applying the residual bottleneck technique. Therefore it uses a 2D-convolutional with 1×1 filter by reducing the channel at the beginning of the operation layer. It is followed by batch normalization (BN) and ReLU activation (δ), defined as:

$$RC_i = \delta(C_{1+BN}(\delta(C^*_{1+BN}(x_i)))) + x_i. \qquad (3)$$

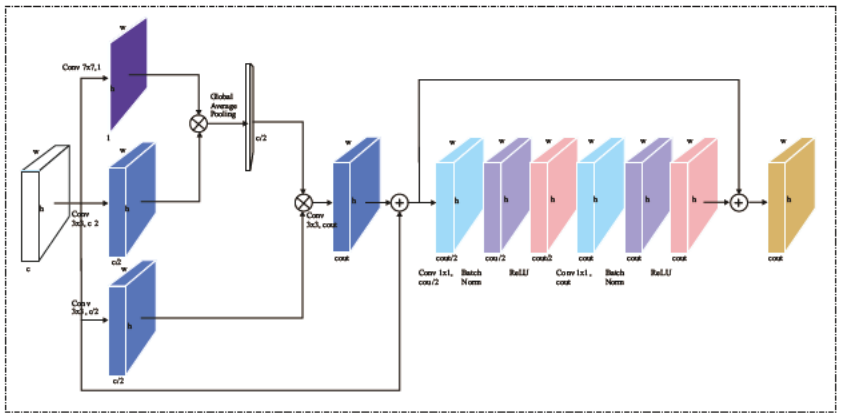

Fig. 2.: A represented global context module captures the specific features.

It creates a smaller channel in the initial convolution layer C^*_{1+BN}, which pushes the network to use fewer parameters while suppressing excessive computation. The proposed attention module effectively separates important features from trivial information. In this case, the top part of the facial area, such as the eyes, eyebrows, and forehead, is essential information to encourage the classification system to predict emotion categories accurately. In addition, this module can enhance the correct linkage of facial elements, which prevents the prediction system from generating large false positives. To minimize the error that causes over convolutional process, it applies a skip connection that sums the output block with input features.

2.3 Classifier Module

A classification system requires a classifier block to predict the output of a set of categories. This module is generally placed at the network's end after passing through several feature extractors. The classifier module used applies global average pooling after feature extraction to generate summary vectors from each feature set. It then applies a neural network layer to generate a logit score with the number of vector dimensions corresponding to the facial expression category. At the end of the module, a softmax activation is employed to form a normalized probability from the vectorized logit scores. The result of this process contains the prediction weights of each facial expression class, with the maximum probability being the final prediction of the proposed system.

2.4 Integrated Module

The proposed masked facial expression offers a real-time application that can estimate facial emotions from predefined localization regions. Therefore, the integrated system process applies face detection to distinguish faces and other objects at the beginning of the process. Our system uses ACETRON [13] as a reliable face detector in the machine and user interaction application. This stage produces a cropped face region of interest to be used as input for the classification system at the inference stage.

3. DATASETS AND SETTINGS

3.1 Dataset and Augmentation

This study created a masked facial expression dataset that modifies the Karolinska Directed Emotional Faces (KDEF) [14] dataset to obtain knowledge of masked faces, as shown in Fig 3 (a). It applies mask manipulation to the entire facial image, covering the nose to chin region. This modification implements [15], which applies face detection to obtain the face region and then a landmark approach to determine the position and location of the mask. The dataset consists of 4900 images of human facial emotion, with 70 individuals showing seven primary expression faces (happy, angry, neutral, fear, disgust, surprise, and sad). Each class provides a variety of facial poses, such as straight, full left profile, full right profile, half left profile, and half right profile. The dataset we produce is a cropped region of the face that can encourage the prediction model performance. It also resizes the cropped RGB image with 48×48. This study uses the augmentation method to increase the variety of data by applying lighting and color distortion. Besides, it uses rotation and horizontal flips to enrich the knowledge of facial poses.

On the other hand, training and testing of the model were also performed on the M-LFW dataset provided by [16]. This model creates masked faces from the Labeled Faces in the Wild (LFW) dataset, as shown in Fig 3 (b). This dataset consists of three types of facial expressions (positive, negative, neutral), each category of which contains five face orientations (center, left, right, top, and bottom). The total images used are 13000 examples selected from the LFW dataset. The proposed model uses cropped face images and applies augmentation techniques to the entire image. To homogenize the training knowledge, we use the same augmentation approach applied to M-KDEF.

3.2 Training configuration

The proposed classification system applies several configurations to optimize the training process. It uses the Categorical cross-entropy loss function to calculate the difference between prediction and ground truth. In

Fig. 3.: M-KDEF dataset (a), and M-LFW dataset (b)

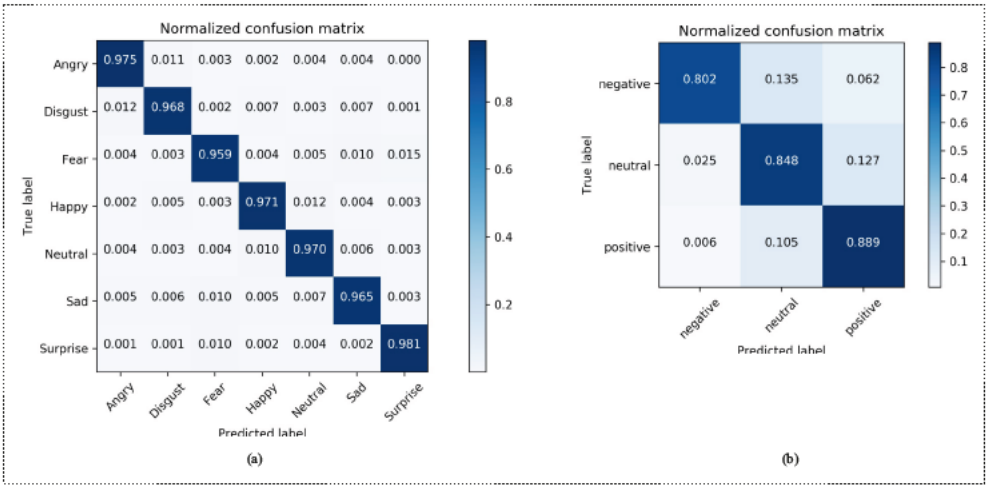

Fig. 4.: Confusion matrix on M-KDEF (a) and Confusion matrix on M-LWF (b).

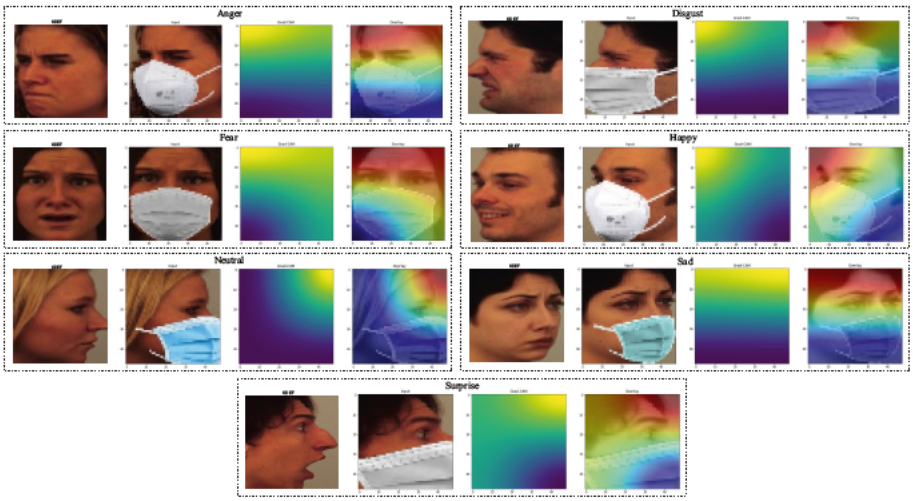

Fig. 5.: The class activation map observes the essential facial components evaluated on the M-KDEF dataset. It investigates feature maps at the end layer of the proposed backbone.

the training process, the proposed model uses Adaptive Moment Estimation (Adam) with an epsilon of 10^{-7} and a starting learning rate of 10^{-4}. The update process is performed on the learning rate by multiplying 0.75 when the training performance does not increase in 20 epochs. Training and testing simulations were conducted on the Keras framework. The described configuration was applied to M-KDEF and M-LFW.

In the KDEF dataset, the training stage uses a batch size of 128. Initial training of the model is performed on the original KDEF in the 50th epoch at 10-fold. And then continue to train the model in the M-KDEF dataset. We use 10-fold cross-validation to split and evaluate the model in this dataset.

Table 1.: Evaluation of the proposed model on M-LFW and M-KEF datasets

Dataset	Model	Accuracy (%)	GFLOPS	Parameters
M-LFW	VGG19	65.42	19.7G	144M
	MobileNetV2	67.18	0.3G	3.5M
	ResNet50	69.92	25.7G	4.3M
	ViT [17]	76.23	17.6G	86.7M
	RAN [18]	79.20	15.3G	12.5M
	ACNN [19]	82.53	18.1G	157M
	OADN [20]	84.21	11.5G	87M
	Yang et al 2021 [8]	87.92	-	-
	Yang et al 2022 [21]	90.31	5.3G	23.5M
	Proposed Slim + RGC	**87.14**	**0.019G**	**756K**
M-KDEF	MobilenetV1	95.48	0.049G	3.2M
	MobilenetV2	93.93	0.03G	2.3M
	MobileNetV3	93.03	0.04G	5.1M
	ShuffleNetV1	88.30	0.015G	973K
	ShufflenetV2	90.74	0.055G	4.0M
	VGG16	53.53	1.41G	14.7M
	ResNet18	82.63	0.065G	11.2M
	GhostNet	94.84	0.02G	3.9M
	Proposed Slim	**94.90**	**0.016G**	**376K**
	Proposed Slim + RGC	**96.99**	**0.019G**	**756K**

Table 2.: Comparison of runtime efficiency

Model	ACC (%) on M-KDEF	Parameters	GFLOPS	FPS classication	FPS integrated
MobilenetV1	95.48	3,236,039	0.04871	168.15	57.19
MobilenetV2	93.93	2,266,951	0.03391	124.64	51.72
MobileNetV3	93.03	5,127,839	0.04045	110.38	48.51
ShuffleNetV1	88.30	973,567	0.01531	192.20	60.15
ShufflenetV2	90.74	4,025,915	0.05459	109.54	48.19
VGG16	53.53	14,718,279	1.41	78.04	41.69
ResNet18	82.63	11,198,919	0.06506	123.35	51.08
GhostNet	94.84	3,918,680	0.0215	123.12	50.60
Proposed Model	**96.99**	**756,447**	**0.01914**	**203.73**	**61.30**

4. EXPERIMENTAL RESULTS

This section evaluates the slim architecture on masked facial expression datasets. Additionally, it examined the processing data speed, which was tested on a CPU device and compared to other architectures.

4.1 Ablative Study

The proposed masked facial emotion recognition uses a slim backbone that efficiently extracts distinctive features. This module helps the whole system operate at real-time speed without significantly reducing accuracy. Experimental results show that this module achieves 94.40% on the M-KDEF dataset with 376K parameters. It also generates a small computational power, resulting in 0.016 GFLOPS. The following observation shows an improvement in accuracy when implementing the Represented Global Context Module. It increases the accuracy by 2.09% on the M-KDEF dataset while increasing the parameters by 380K. In addition, the impact of using this module is a slight increase in computational usage of 3MFLOPS. The complete experiment is shown in Table 1.

The proposed backbone was also investigated comprehensively to determine the influential facial features for classifying masked facial emotions. This evaluation evaluates the determination of essential elements by applying the Grad-CAM technique to find a heat map. Fig 5 shows that our model avoids the mask component and focuses on the upper face area. The eyes, eyebrows, and Forehead are the features of concern from the proposed network.

4.2 Evaluation on Datasets

4.2.1 M-LFW

This dataset represents masked faces taken from an LFW face recognition dataset. This mask dataset provides fewer categories than a typical facial expression dataset. The proposed model was trained and evaluated on this dataset and compared its performance with previous models and work. Our model achieved 87.34%, which outperformed ACNN and other methods below it, as shown in Table 1. However, the accuracy of the proposed model is lower than Yang et al. [21] as the leading

Fig. 6.: System integration testing on real-world scenarios in 640 x 480 resolution using a webcam.

competitor. It differs by 3.18 %, but our model significantly produces lighter parameters and computation. The confusion matrix in Fig. 4 (b) presents that the positive category obtained a higher true positive score than the other classes. On the other hand, our model achieves a prediction error of 13.5% when predicting negative as the neutral class.

4.2.2 M-KDEF

This dataset is a modification of the KDEF dataset by implementing a mask on the face area. It also provides seven basic facial emotion categories commonly used in facial expression datasets. We applied 10-fold to the training process using cross-validation to create a fair evaluation. The experimental results show that the proposed module achieves 96.99% on the dataset outperforming MobileNetV1 and other models. Table 1 shows that the proposed module obtains the highest performance of the Benchmark mobile model. The proposed model also has the added value of high-cost efficiency. The confusion matrix in Fig. 4 (a) shows that the surprise category has a higher true positive value than other expression classes. On the other hand, our model achieves the highest false prediction of 1% when predicting fear as sadness and surprise classes.

4.3 Runtime Performance on CPU

Model speed testing obtains the capability of a method in practical application issues. This observation highlights the efficiency value by focusing on implementing a low-cost device. Furthermore, the application issue weakens the deep learning model that tends to operate slowly on low-cost devices reducing the method's applicability to real-world cases. The proposed model has evaluated the efficiency by measuring the CPU speed on a PC Desktop using Intel Core i7-6700T CPU @2.80GHz with 16 GB RAM. The results shown in Table II demonstrate that the proposed classification model can operate fast on this device by 203.73 FPS. Moreover, it is supported by a high-efficiency level with a number of parameters and computational complexity of 756,447 and 0.019, respectively. The proposed model is integrated with an ACETRON face detector to measure its capabilities in real-world scenarios. We install the classification model at the end of the whole model after acquiring the face regions generated by face detection. This integrated model obtained a speed of 61.30 FPS which is faster than the integrated mobile benchmark model. It differs by 1 FPS from ShuffleNetV1, which is slower than the proposed recognition system. Qualitative results in Fig 6 show that the proposed system can recognize facial expressions even when covered by a mask. This simulation uses the M-KDEF knowledge, which has data on pose variation and balance in the number of each class.

The proposed deep learning model can identify facial expressions covered by masks by focusing on the upper part of the face. Based on the ablation study, our model focuses on the eye, eyebrow, and chin areas to recognize facial emotions. The slim backbone module and efficient attention blocks satisfactorily extract these features resulting in accurate predictions. On the other hand, the proposed model does not impose a significant burden on parameters and computation. Thus it builds a lightweight model and friendly network on low-cost devices.

5. CONCLUSION

This paper presents a lightweight architecture to recognize masked facial expressions using the CNN method. It focuses on improving the accuracy and efficiency of the implementation on a CPU device. The proposed architecture has a slim backbone and represents a global context module. Inverted residual is used in the backbone to extract spatial information rapidly. While represented global context is offered to capture the important feature in a wide range of relationships. The masked face encourages our model to decide the critical information is the upper face. As a result, the proposed model classification achieves a competitive accuracy when compared with the previous method and the mobile benchmark model. Additionally, the integrated model with face detection obtains the fastest processing data speed compared to mobile architecture, which achieves 61.30 FPS on an Intel Core i7-6700T CPU. A combination of loss functions can improve prediction performance without reducing its efficiency, which is the further work of this study.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the government(MSIT).(No.2020R1A2C200897212).

REFERENCES

[1] Y. Xia, H. Yu, X. Wang, M. Jian, and F.-Y. Wang, "Relation-aware facial expression recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 1143–1154, 2022.

[2] M. D. Putro, D.-L. Nguyen, and K.-H. Jo, "Real-time multi-view face mask detector on edge device for supporting service robots in the covid-19 pandemic," in *Intelligent Information and Database Systems*, N. T. Nguyen, S. Chittayasothorn, D. Niyato, and B. Trawiński, Eds. Cham: Springer International Publishing, 2021, pp. 507–517.

[3] J. Yang, T. Qian, F. Zhang, and S. U. Khan, "Real-time facial expression recognition based on edge computing," *IEEE Access*, vol. 9, pp. 76 178–76 190, 2021.

[4] M. Rizzato, M. Antonelli, S. D'Anzi, C. Di Dio, A. Marchetti, and D. Donelli, "The impact of face masks used for covid-19 prevention on emotion recognition in facial expressions: An experimental study," *Biology and Life Sciences Forum*, vol. 19, no. 1, 2022.

[5] A. M. F. A. Nawal Younis Abdullah, "Masked face with facial expression recognition based on deep learning," *The Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 1, 2022.

[6] M. Gori, L. Schiatti, and M. B. Amadeo, "Masking emotions: Face masks impair how we read emotions," *Frontiers in Psychology*, vol. 12, 2021.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[8] B. Yang, W. Jianming, and G. Hattori, "Face mask aware robust facial expression recognition during the covid-19 pandemic," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 240–244.

[9] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[10] M. D. Putro, D.-L. Nguyen, and K.-H. Jo, "A fast cpu real-time facial expression detector using sequential attention network for human–robot interaction," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 11, pp. 7665–7674, 2022.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[13] M. D. Putro, A. Priadana, D.-L. Nguyen, and K.-H. Jo, "A faster real-time face detector support smart digital advertising on low-cost computing device," in *2022 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, 2022, pp. 171–178.

[14] M. Calvo and D. Lundqvist, "Facial expressions of emotion (kdef): Identification under different display-duration conditions," in *Behav. Research Methods*, vol. 40, no. 2008, 1998, p. 109–115.

[15] A. Anwar and A. Raychowdhury, "Masked face recognition for secure authentication," 2020.

[16] B. Yang, W. Jianming, and G. Hattori, "Face mask aware robust facial expression recognition during the covid-19 pandemic," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 240–244.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[18] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.

[19] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019.

[20] H. Ding, P. Zhou, and R. Chellappa, "Occlusion-adaptive deep network for robust facial expression recognition," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 2020, pp. 1–9.

[21] B. Yang, J. Wu, K. Ikeda, G. Hattori, M. Sugano, Y. Iwasawa, and Y. Matsuo, "Face-mask-aware facial expression recognition based on face parsing and vision transformer," *Pattern Recognition Letters*, vol. 164, pp. 173–182, 2022.