Low-resolution facial emotion recognition on low-cost devices

Muhamad Dwisnanto Putro, Jane Litouw, Vecky Canisius Poekoel
Department of Electrical Engineering, Faculty of Engineering, Sam Ratulangi University, Manado, Indonesia

Article Info

Article history:

Received Jul 31, 2023
Revised Dec 8, 2023
Accepted Jan 6, 2024

Keywords:

Deep learning
Facial emotion recognition
Low resolution
Low-cost devices
Real-time model

ABSTRACT

The low-resolution input image is a crucial challenge for applying facial emotion recognition in real-world scenarios. The critical problem is that valuable object features are relatively lost in the extraction process due to their small size. On the other hand, this vision system is required by a machine to run smoothly on low-cost devices. Facial emotion recognition using a lightweight feature extractor is proposed in this study to effectively capture crucial facial components in a low-resolution image. To compromise the running speed, this work offers an efficient feature convolution to discriminate specific facial features. In addition, the system is embedded with an attentive module to capture important features and correlate them. Our model performance is evaluated on low-resolution public datasets achieving the accuracy of 97.34%, 81.10%, and 80.12% on Karolinska directed emotional faces (KDEF), real-world affective faces database (RFDB), and facial expression recognition 2013 plus (FER2013Plus), respectively. The practical application demands that the deep learning model can operate fast on inexpensive devices. Consequently, the model achieved a speed of 290 frames per second (FPS) on a central processing unit (CPU) device.

Corresponding Author:

Muhamad Dwisnanto Putro
Department of Electrical Engineering, Faculty of Engineering, Sam Ratulangi University
Kampus Bahu UNSRAT, Manado, Indonesia
Email: dwisnantoputro@unsrat.ac.id

1. INTRODUCTION

Facial emotion recognition is an approach to identifying a person's expression based on facial gestures. Computer vision commonly performs this work by recognizing facial features that influence an expression [1]. A face has attributes that cooperate to construct a gesture. So the relationship between the element shapes is the distinctive information of each facial emotion. The nose, Mouth, Eyes, Eyebrows, and Cheeks are essential objects of the face related to facial expressions [2]. Therefore, facial emotions present a correlation model between these important facial components. Several conventional methods struggle to correlate the features precisely, resulting in low performance [3]. The real application demands that this approach operate accurately and efficiently, especially in robotics. Social robots require interacting with humans at all times, thus necessitating a facial emotion recognition system as a non-verbal sensor. It expresses the user response showing liking or disliking in communication activity. So, this approach is needed by a robot to predict the person's interest or reaction, which can affect the machine's response correctly. The cognitive prediction error affects human-machine interaction failure, as well as misunderstanding the interpretations [4]. The challenges of real case scenarios present a vision system to be robust without being compromised by disturbance [5]. This system's performance drops when faced with low-resolution input images. These cases are common and involve long-distance human objects with small faces. Facial expression recognition must recognize and detect human faces at various ranges. So this challenge becomes a crucial issue for facial emotion recognition in real-world applications. Low-quality images make it difficult to capture some parts of essential features. Eye and eyebrow objects tend to be lost during the extraction process when applying deep learning methods, even though these features are essential information for each facial expression. Therefore, an extraction method that can work optimally for low-resolution images is a priority concern in this work.

Deep learning has been present as an effective feature learning method to separate essential elements from the background. Convolutional neural network (CNN) is a widely used approach to capture object features [6]-[8]. It employs a learnable weighted filter operating on the spatial area of the image that effectively filters out the target object. Facial expression classification systems have obtained high accuracy by applying this method [9]-[11]. Combining convolution modules distinguishes specific facial features and increases their intensity to encourage accurate prediction results in the classifier. Nevertheless, the limited spatial area reduces the performance of CNN to associate features separated by a long-range distance. Previous works applied attention modules to overcome this problem [12]-[14]. Furthermore, recent works utilize CNN to predict small facial expressions with competing accuracy [15]-[19]. Deep learning relates small facial features that are lost due to the over-extraction process. However, these works did not conclusively prove that their models can operate effectively in real-world scenarios. In addition, the classic issue of deep learning models is the dependence on high computation and the abundant number of parameters [20]. This problem reduces the data processing speed of benchmark models on low-cost devices. The work in this article proposes a novel, efficient deep learning-based model to predict facial emotions in low-resolution images. An efficient feature convolutional (EFC) is introduced to filter essential facial features related to expression without requiring excessive computation and parameters. This network establishes a real-time human facial emotion system and is tested on real-case scenarios to observe the effective performance of the model implementation. We summarize the contribution of the work as follows:

- A novel, efficient feature extractor is proposed using a lightweight convolution operation that effectively distinguishes important facial components from trivial features. It applies a combination of standard convolution and depth-wise convolution layers to reduce the parameter model and computational complexity.
- A simple attention module is offered to improve the performance of lightweight convolutional blocks that are easy to plug and play. It comprehensively enhances specific facial information by correlating low-intensity crucial features.
- A facial emotion recognition system integrated with lightweight face detection presents a robust system that can be rapidly and reliably implemented in real applications.
- This work comprehensively evaluates the model accuracy on low-quality benchmark datasets. It also assesses the model's speed when implemented on low-cost devices operating effectively in real-world scenarios. The proposed model can achieve fast speed processing data of 290 frames per second (FPS) that is implemented on a low central processing unit (CPU) frequency.

2. RELATED WORKS

A study has applied the conventional method [16] to classify low-resolution facial expressions. It uses the LBP method to acquire facial gesture features and applies support vector machine (SVM) to classify the results. This approach was tested on Cohn-Kanade and PETS 2003 datasets, achieving low accuracy in predicting multi-pose faces. Yan *et al.* [19] introduces a filter learning model for low-resolution images formulated using linear operations. Linear discriminant analysis (LDA) optimizes the image filter learning process. The method was only tested on a formal dataset that does not represent implementation in a real case. A deep learning network shows satisfying results compared to conventional methods for predicting facial expression categories in low-resolution images. Lo *et al.* [15] has proposed an uncertainty modeling to separate facial expression categories using the CNN model. A combination of five loss types was used to evaluate the model's performance by providing a penalty for correct and incorrect prediction results. In addition, it implements a ten-layer feature extractor, thus claiming that the resulting model can operate in real-time. Using multiple hyperparameters causes this work to depend on the tuning process, which is expensive. Furthermore, Bodavarapu and Srinivas [17] has designed a novel CNN model (FERConvNet) with small dimensional inputs. It employs a 2D-convolutional layer followed by batch normalization and dropout to inhibit overfitting problems.

This network was performed superior to visual geometry group (VGG) architectures on the facial expression recognition 2013 (FER2013). The runtime efficiency shows that this model is promising for real-world applications. On the other hand, a generative adversarial network with a super-resolution method has been proposed to predict low-resolution facial expressions [18]. This work presents four crucial parts of the model: backbone, feature generator, classifier, and discriminator. The applied upsampling and twin backbone methods slow down the data processing speed of this method.

3. PROPOSED ARCHITECTURE

A deep learning method employs a feature extractor based on data learning to distinguish the interesting facial elements that can encourage the classifier to predict accurately. The proposed architecture adopts this approach that implements two core modules, such as a backbone and a classifier, as presented in Figure 1. The backbone plays a crucial role as a feature distinguisher that implements lightweight convolution operations by applying efficient convolutional features (ECFs). And then, the classifier focuses on generating the final prediction of facial emotion.

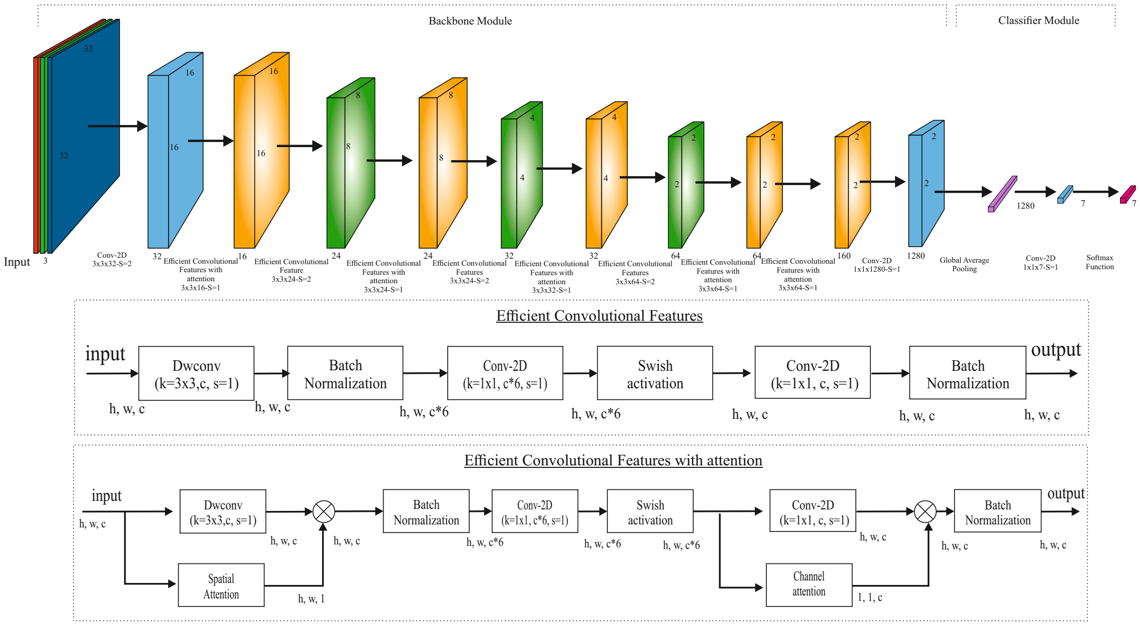


Figure 1. The proposed overall network that consists of two core modules. A lightweight backbone comprehensively extracts the facial features. Then, a simple classifier predicts the emotional classes

3.1. Backbone module

Modern classification networks apply a backbone as a core block to capture the distinctive object. It sequentially involves convolutional operations followed by normalization and activation [21]. The proposed architecture utilizes lightweight convolution operations to compress the computational complexity and number of parameters. It will automatically lead to increased data processing speed for implementation in real applications. Standard convolutional with the 3×3 kernel is utilized at the beginning of the network to shrink the map dimension. On the other hand, the end of the network involves the 1×1 filter to enrich the feature selection.

3.1.1. Efficient convolutional features module

The proposed architecture employs a lightweight convolutional module incorporating sparse kernel operations to reduce redundancy and balance features and parameters. An ECF is presented as a combined convolutional module delivering higher data processing speed than residual convolutional blocks [7]. Figure 1 in the middle part shows that it utilizes depth-wise convolutional blocks at the beginning of the process to extract features in the spatial regions operated on each channel. It applies batch normalization after the depth-wise convolutional operation to control the variety of feature map information and dramatically reduce the overfitting problem. The channel-based information blending process is performed by 2D-convolutional that applies the 1×1 filter followed by improved Swish activation [22]. This block increases the number of feature map channels to generate varied features. Furthermore, a simple convolution operation is also applied at the end of this module by downsizing the number of channels which is the same size as an initial channel, an explanation of the ECF module is formulated as follows:

$$EFC(x) = \mathbb{N}(conv_{1x1}(conv_s(\mathbb{N}(DW(x))))), \tag{1}$$

where

$$conv_s(x) = \delta(conv_{1x1}(x)). \tag{2}$$

The input feature x goes through a sequential depth-wise convolutional (DW) and 2D-convolutional with 1×1 filter ($conv_{1x1}$) followed by batch normalization ($\mathbb{N}$) to prevent overfitting in the training phase. It employs smooth Swish activation to preserve the negative score by applying the effect of a smooth continuous function, (δ). S-Swish activation is proposed to activate the neurons of the convolution output by smoothly adjusting the negative values. The value of β is a constant parameter to set the smoothing magnitude on the negative region. It will provide a more extensive scope of negative scores if the parameter is small and vice versa. The proposed activation is formulated as follows:

$$\delta(x) = \frac{x}{1 + exp^{-\beta x}}. \tag{3}$$

The proposed architecture utilizes the EFC module to quickly reduce the feature map without compromising computational speed and avoiding parameter abundance. The general deep learning architecture implements a deep convolution layer that produces low efficiency. The proposed EFC module tackles this issue by implementing lightweight convolution operations. The depth-wise operation employs a weighted kernel utilizing only each channel to save parameters. On the other hand, the simple convolution operation works by finding the feature relations of all channels at the same position, which can increase the effectiveness of the proposed module.

3.1.2. Efficient convolutional features with the attention module

The attention module performs satisfactorily in deep learning networks because it can improve accuracy without significantly decreasing model efficiency. This method works like the human eye, which focuses on sharp vision to find influential features of the target object. The specific facial areas for recognizing expressions are usually related to the mouth, eyes, and nose. The attention module can capture the target object's specific features through weighted learning. It can also highlight the features of interest and reduce the trivial components. In addition, it can strengthen the intensity of the relationship between essential facial elements and reduce the correlation of components unrelated to the prediction. An attention module is usually employed in a deep layer to improve saturation accuracy. However, this work applies it in a shallow layer to sustain the performance of the lightweight convolutional layer. The proposed network applies two attentive modules to improve the ECFs module on specific network parts. Each attention module is assigned to the feature map's spatial operation and channel reconstruction block. This fusion module can increase the effectiveness of the EFC module in extracting features without reducing the map dimensions. The integrated model is illustrated as

$$EFC(x) = \mathbb{N}(conv_{ch}(conv_s(\mathbb{N}(DW_{sp}(x))))), \tag{4}$$

where

$$conv_s(x) = \delta(conv_{1\times 11}(x)), \tag{5}$$

$$DW_{sp}(x) = DW(x) \otimes SP_{att}(x), \tag{6}$$

$$conv_{ch}(x) = conv_{1\times 1}(x) \otimes CH_{att}(x). \tag{7}$$

An input feature x is initially extracted by the spatial enhancement operation DW_{sp}, followed by $conv_{ch}$ as a channel extraction of a single spatial feature. A spatial attention module is embedded at the beginning of the process of the ECF module to boost the extraction feature from spatial feature areas. It updates the output of the depth-wise convolutional operation, as shown in Figure 2(a). This attention module applies a simple convolutional with a 1×1 filter to generate a single-channel feature map representation. The spatial attention module can be formulated as

$$SP_{att}(x) = \zeta(conv_{1\times1}(x)), \tag{8}$$

ζ is a sigmoid activation used to convert integer values to weighted probabilities, and $conv_{1\times1}$ is a 2D-convolutional operation with a 1×1 filter using a single channel. Furthermore, the last block part of an EFC embeds an attentive channel module that can capture the feature representations from each channel. It employs a global average pooling (GAP) to find the channel-based characterization from the average operation on each map, as illustrated in Figure 2(b). This channel attention module can be formulated as

$$CH_{att}(x) = \zeta(conv_{1\times1}(GAP(x))), \tag{9}$$

where $conv_{1\times1}$ is a 1×1 convolution and ζ is a sigmoid function. The applied attention module emphasizes improving the model's accuracy by filtering specific proposed efficient operations. The spatial attention module captures positional neighborliness features containing important facial elements. Meanwhile, the channel attention module connects long-range features that can enhance specific facial features. This integration module effectively boosts the whole network's performance to predict seven facial emotions.

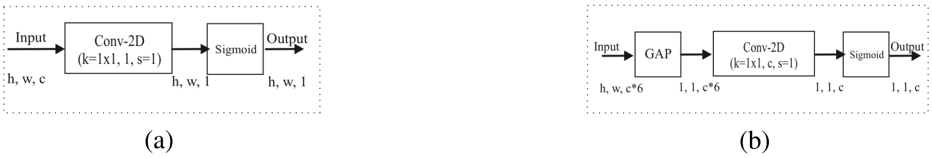


Figure 2. The attention module is used in an ECFs block to enhance performance: (a) a spatial attention module and (b) a channel attention module

3.2. Classifier module

Image recognition of deep learning usually uses a classifier module in the last network to predict the label categories. It applies the fully connected block to produce the vector according to the number of classes. This operation relates all the vector features and models the connection through the trained weights. The proposed network applies 2D convolution with a 1×1 filter to generate 1280 features, which is also the number of channels. This layer is employed at the end of the EFC module. Then, a global average pooling is utilized to summarize the represented features by finding the average of each map to avoid redundant parameters. A 2D-convolutional is also employed to create a vector that has a dimensional size equal to the number of predicted emotion classes. Furthermore, a Softmax activation is applied to produce the probabilities associated with a multimodal distribution. This function is applied to tackle multiple facial expression recognition issues.

3.3. Implementation setup

Deep learning networks require a suitable configuration to encourage the training and testing process to run optimally. The hyperparameter setting in the training process adopts the previous research [4] to prevent a vanishing gradient. Model simulation is conducted on Python framework with Keras library for data learning process on Ubuntu operating system. The training stage uses an intel core i7-6700T CPU @2.80GHz processor, 16 GB RAM, and an NVIDIA Titan RTX graphics card. The facial expression network is trained and tested on low-resolution datasets, including Karolinska directed emotional faces (KDEF) [23], real-world affective faces database (RFDB) [24], and facial expression recognition 2013 plus (FER2013Plus) [25]. The augmentation technique is employed only on the KDEF dataset that applies color, brightness, contrast, rotation, and flip transformations. It applies a linear interpolation approach to the whole dataset to generate a small scale of the whole image. The proposed model learns the image of all datasets without pre-trained knowledge. The model utilizes a categorical cross-entropy loss to calculate prediction error, which compares with the ground truth label. The training phase in the KDEF dataset applies a batch size of 128 with 10-fold cross-validation to evaluate the model and split the datasets, where each fold is trained at 50 epochs. The model was also trained and evaluated on RFDB and FER2013Plus by applying a data split configuration that refers [15]. The model was trained on the FER2013Plus dataset in 500 epochs and 32 batch sizes. Overall training process uses an initial learning rate of 10^{-4} with adaptive moment estimation (Adam) as the optimizer. The updating learning rate is performed by multiplying 0.75 when the training accuracy does not improve in 20 epochs.

4. EXPERIMENTS AND RESULTS

This section evaluates the proposed model on several benchmark low-resolution FER datasets. It also compares the performance with the previous works. The following subsection discusses the efficiency of modules embedded in low-cost devices.

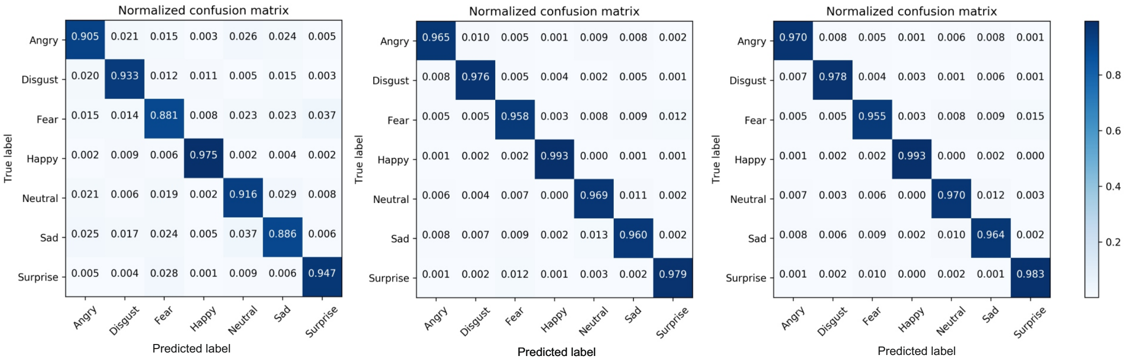
4.1. Evaluation on KDEF dataset

This public dataset provides 4,900 red green blue (RGB) images based on a laboratory environment containing seven basic facial emotions: fear, anger, neutral, sadness, disgust, surprise, and happiness. The original dataset assigns 70 persons to design five poses, such as straight, full-right, full-left, half-left, and half-right poses. It accommodates male and female gender to increase the variety of human faces. We evaluate the proposed low-resolution model in various small-scale images, including 32×32, 20×20, and 10×10. The number of parameters and GFLOPS is generated on 32×32 resolution. We cannot show the competitor's performance at 20×20 and 10×10 resolutions due to feature map dimensional adjustments. Table 1 shows that the proposed model achieves 97.34% accuracy. This result does not outperform heavy architectures such as VGG11 and VGG13. However, the proposed model is superior to ResNet18.

Table 1. Performance results on KDEF datasets in low-resolution images

Model	Parameter	GFLOPS	Accuracy (%)		
			32x32	20x20	10x10
MobileNetV2	2,266,951	0.013	96.59	-	-
MobileNetV1	3,236,039	0.023	96.15	-	-
MobileNetV3 Small	2,949,663	0.017	95.58	94.73	84.64
MobileNetV3 Large	5,127,839	0.018	96.75	95.66	87.32
ShuffleNetV1	973,567	0.006	90.08	-	-
ShuffleNetV2	4,025,915	0.020	96.14	-	-
VGG11	28,137,607	0.344	99.23	-	-
VGG13	28,322,119	0.495	99.30	-	-
ResNet18	11,198,919	0.035	97.12	96.36	91.97
GhostNet	3,918,680	0.012	96.79	96.13	90.11
Proposed	513,484	0.007	97.34	97.13	92.04

On the other hand, the comparison shows that our model outperforms the benchmark mobile model's performance and produces fewer parameters. Although the computational complexity of the proposed model is slightly larger than ShuffleNetV1, our model obtains better accuracy than the model. An extensive investigation was conducted on the performance of the classification system shown on confusion matrices in Figure 3. The happy category obtains the best true positives on all low-resolution images. In contrast, the fear expression obtained the highest false positives on all resolutions tested.



(32x 32 pixels) (20 x 20 pixels) (10 x 10 pixels)

Figure 3. Confusion matrix of the proposed model evaluated on KDEF dataset in 32×32, 20×20, and 10×10 resolutions

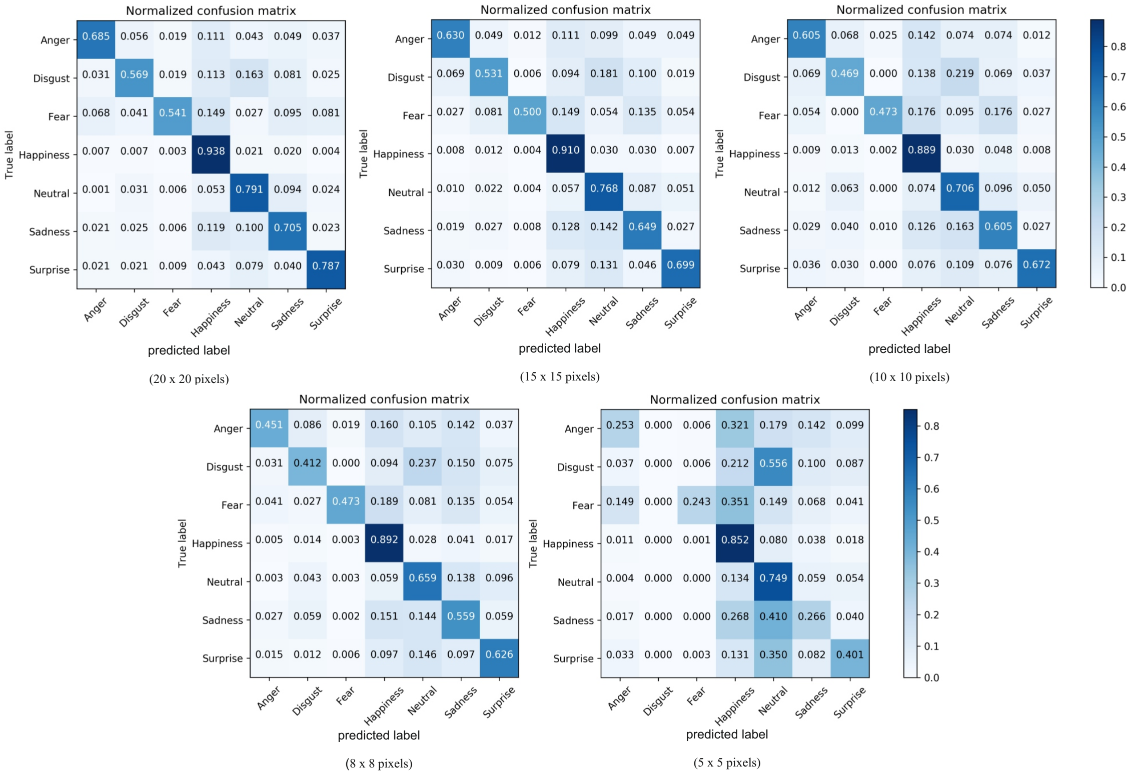
4.2. Evaluation on real-world affective faces database dataset

This wild dataset consists of 30,000 facial images that contain various poses. The images in this dataset are taken from an unconstrained environment, with natural facial gestures. The official website provides single-expression and multi-expression labels. However, this work uses single labels to evaluate facial expressions in low-resolution images. Our experiments use seven classes of standard emotions, such as disgust, surprise, fear, neutral, happiness, anger, and sadness. Table 2 presents our model achieves superior performance to other competitors at 10×10, 8×8, and 5×5 pixels. It even outperforms the state-of-the-art work [15]. However, this network performs poorly than these competitors at 15×15 resolution, which differs by 0.39% and achieves similar accuracy at 20×20 pixels.

Table 2. Proposed model compared to other methods on RFDB datasets in low-quality images

Model	Accuracy (%)				
	20×20	15×15	10×10	8×8	5×5
Lownet [26]	0.7070	0.6897	0.6473	0.6111	0.5551
APEN [27]	0.7792	0.7362	0.6865	0.6481	0.5747
SCN [28]	0.6926	0.5613	0.5555	0.4492	0.4182
DMUE [29]	0.7363	0.706	0.6213	0.5593	0.4654
RUL [30]	0.8063	0.7565	0.6917	0.6406	0.5616
MULR [15]	0.8110	0.7744	0.7096	0.6630	0.5918
Proposed	0.8110	0.7705	0.7344	0.7011	0.5991

The investigation of the prediction of each category is shown in the confusion matrix. Figure 4 shows that happiness has the highest true positive value compared to the other six facial emotions. This expression shows a unique facial gesture compared to the others. Fear emotion gets the highest false positive when evaluating it on pixels 20×20, 15×15, and 8×8. In comparison, the disgust emotion reaches the lowest accuracy on resolutions 8×8 and 5×5. Even the model fails to recognize this expression on the smallest pixels.



	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	0.685	0.056	0.019	0.111	0.043	0.049	0.037
Disgust	0.031	0.569	0.019	0.113	0.163	0.081	0.025
Fear	0.068	0.041	0.541	0.149	0.027	0.095	0.081
Happiness	0.007	0.007	0.003	0.938	0.021	0.020	0.004
Neutral	0.001	0.031	0.006	0.053	0.791	0.094	0.024
Sadness	0.021	0.025	0.006	0.119	0.100	0.705	0.023
Surprise	0.021	0.021	0.009	0.043	0.079	0.040	0.787

(20 x 20 pixels)

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	0.630	0.049	0.012	0.111	0.099	0.049	0.049
Disgust	0.069	0.531	0.006	0.094	0.181	0.100	0.019
Fear	0.027	0.081	0.500	0.149	0.054	0.135	0.054
Happiness	0.008	0.012	0.004	0.910	0.030	0.030	0.007
Neutral	0.010	0.022	0.004	0.057	0.768	0.087	0.051
Sadness	0.019	0.027	0.008	0.128	0.142	0.649	0.027
Surprise	0.030	0.009	0.006	0.079	0.131	0.046	0.699

(15 x 15 pixels)

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	0.605	0.068	0.025	0.142	0.074	0.074	0.012
Disgust	0.069	0.469	0.000	0.138	0.219	0.069	0.037
Fear	0.054	0.000	0.473	0.176	0.095	0.176	0.027
Happiness	0.009	0.013	0.002	0.889	0.030	0.048	0.008
Neutral	0.012	0.063	0.000	0.074	0.706	0.096	0.050
Sadness	0.029	0.040	0.010	0.126	0.163	0.605	0.027
Surprise	0.036	0.030	0.000	0.076	0.109	0.076	0.672

(10 x 10 pixels)

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	0.451	0.086	0.019	0.160	0.105	0.142	0.037
Disgust	0.031	0.412	0.000	0.094	0.237	0.150	0.075
Fear	0.041	0.027	0.473	0.189	0.081	0.135	0.054
Happiness	0.005	0.014	0.003	0.892	0.028	0.041	0.017
Neutral	0.003	0.043	0.003	0.059	0.659	0.138	0.096
Sadness	0.027	0.059	0.002	0.151	0.144	0.559	0.059
Surprise	0.015	0.012	0.006	0.097	0.146	0.097	0.626

(8 x 8 pixels)

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	0.253	0.000	0.006	0.321	0.179	0.142	0.099
Disgust	0.037	0.000	0.006	0.212	0.556	0.100	0.087
Fear	0.149	0.000	0.243	0.351	0.149	0.068	0.041
Happiness	0.011	0.000	0.001	0.852	0.080	0.038	0.018
Neutral	0.004	0.000	0.000	0.134	0.749	0.059	0.054
Sadness	0.017	0.000	0.000	0.268	0.410	0.266	0.040
Surprise	0.033	0.000	0.003	0.131	0.350	0.082	0.401

(5 x 5 pixels)

Figure 4. Confusion matrix of the proposed model evaluated on RAFDB dataset in 20×20, 5×15, 10×10, 8×8, and 5×5 resolutions

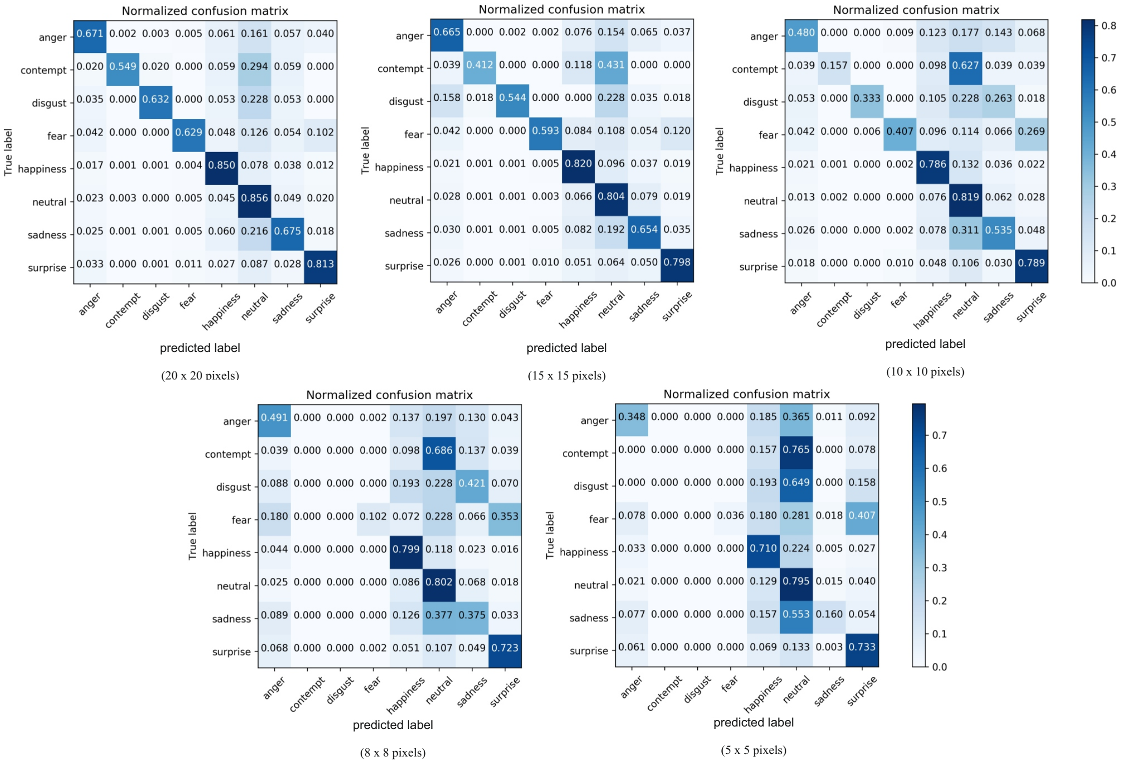
4.3. Evaluation on FER2013Plus dataset

This dataset is a refinement of the original FER2013 dataset, providing better quality ground truth. It has been filtered from non-faces and unknown objects so that some works perform better than the original dataset. It contains eight facial expression categories: contempt, surprise, disgust, fear, neutral, happiness, anger, and sadness. This dataset has been officially divided into the training, validation, and testing datasets. We adopt [29] to merge the training and validation sets used in the training phase. Meanwhile, a test set is utilized to evaluate the model. In this dataset, our model achieves state-of-the-art in all low-quality resolutions. Table 3 illustrates that the model outperforms all competitors, including recent work [15].

Table 3. Proposed model compared to other methods on FER2013Plus datasets in low-quality images

Model	Accuracy (%)				
	20×20	15×15	10×10	8×8	5×5
Lownet [26]	0.6757	0.6589	0.6119	0.5826	0.5190
APEN [27]	0.7587	0.7081	0.6189	0.5896	0.5115
SCN [28]	0.7809	0.7510	0.6962	0.6561	0.5271
DMUE [29]	0.7423	0.6815	0.5957	0.5068	0.4099
RUL [30]	0.7916	0.7580	0.6869	0.6475	0.5823
MULR [15]	0.7957	0.7641	0.7187	0.6752	0.6139
Proposed	0.8012	0.7667	0.7233	0.6829	0.6183

The proposed model also examined the accuracy of each emotion class in this public dataset. Figure 5 presents the confusion matrices that evaluate the predictions of this classification system. The happiness expression has the highest true positive at 20×20 and 15×15 pixels, while the 10×10, 8×8, and 5×5 resolutions indicate that the neutral category has the highest accuracy. Furthermore, this evaluation demonstrates that the most significant prediction error is obtained for the contempt class on all pixel sizes. It describes that the model predicts incorrectly on all test data at 8×8 and 5×5 resolutions. The same problem also occurs with a disgusted expression.



(20 x 20 pixels)

(15 x 15 pixels)

(10 x 10 pixels)

(8 x 8 pixels)

(5 x 5 pixels)

Figure 5. Confusion matrix of the proposed model evaluated on FER2013Plus dataset in 20×20, 15×15, 10×10, 8×8, and 5×5 resolutions

4.4. Runtime efficieny and implementation of real application

The proposed facial emotion recognition system requires face detection to localize the facial area while filtering it from the background. It plays a significant role in focusing the classification model only on the face region, which can improve the adequate performance of the system. Complex backgrounds can reduce the performance of the classification system, so these features need to be discriminated. The proposed system implements fast and accurate face detection on central processing unit (FAFCPU) [4], a fast and accurate face detection for small faces. In the inference stage, the face detection result is cropped and then used as the input of the classification model. Table 4 shows the efficiency results and comparison with other deep learning models. This experiment uses a Logitech c270 webcam as the input stream of the facial emotion recognition system on video graphics array (VGA) resolution (640×480). Face detection is used to filter the face region and improve the effectiveness of the model. A fusion of facial emotion with face detection is measured in integrated speed. The classification model uses a patch image of 32×32 pixels. We tested the speed in 1000 frames and carried the highest value as the measured speed. These measurements were taken on a Jetson Nano and a CPU-based PC to represent a cheap processing device. The proposed model generates lower parameters than the other models. Although the giga floating point operations per second (GFLOPS) of our model is slightly larger than ShuffleNetV1, the proposed model obtains faster data processing speed than this model. It uses cheap operation and avoids deep layers to boost the speed. Our model achieved 47.97 FPS and 290.78 FPS speeds on a Jetson Nano device 4 GB and an Intel Core i7 PC, respectively. In addition, the integrated face detection achieved 20.56 FPS and 68.05 FPS speeds on both devices. The qualitative results of the facial emotion recognition system in low resolution show a satisfying performance. It integrates the proposed model with face detection. The model knowledge data in this test is from the KDEF dataset. Figure 6 presents the system that can recognize facial expressions of small sizes. It even obtains accurate performance on multi-view faces.

Table 4. Runtime efficiency of model compared with benchmark mobile architecture

Model	Parameter	GFLOPS	Acc %	FPS on Jetson Nano		FPS on Intel Core i7-6700T CPU	
				Facial expression	Integrated	Facial expression	Integrated
MobileNetV1	3,236,039	0.023	96.15	19.88	13.24	183.50	56.10
MobileNetV2	2,266,951	0.013	96.59	13.70	10.20	151.60	55.90
MobileNetV3S	2,949,663	0.017	95.58	11.30	8.84	147.05	52.27
MobileNetV3L	5,127,839	0.018	96.75	9.87	7.92	124.74	51.76
ShuffleNetV1	973.567	0.006	90.08	28.26	16.58	187.58	57.19
ShuffleNetV2	4,025,915	0.020	96.14	20.05	13.42	119.08	48.37
VGG11	28,137,607	0.344	99.23	13.75	10.14	89.95	44.34
VGG13	28,322,119	0.495	99.30	12.25	9.30	77.60	41.56
ResNet18	11,198,919	0.035	97.12	8.30	5.75	115.12	49.17
GhostNet	3,918,680	0.011	96.79	18.26	12.24	124.39	48.11
Proposed	513.484	0.007	97.34	42.97	20.56	290.78	68.05

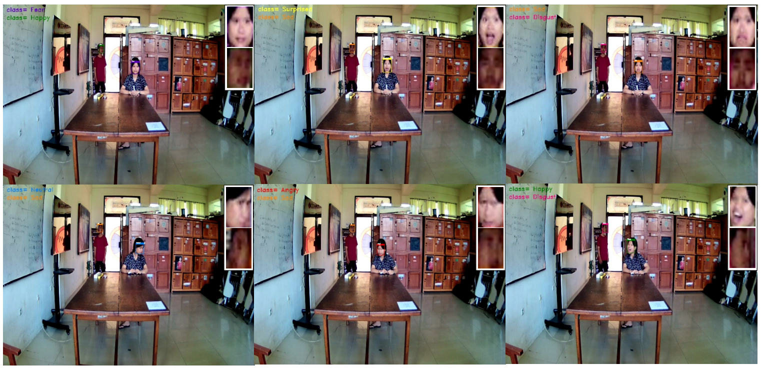


Figure 6. Facial emotion recognition result when integrated the proposed model with face detection. Model predictions are provided on the top left of each image and face patches on top right

5. CONCLUSION

This paper presents an efficient network using deep learning to recognize facial emotion in low-quality images. This work offers ECFs to fast discriminate specific features related to facial gestures. It also applies a simple attention module that enhances the ECF module's performance. It can capture interesting information on channel and positional feature maps. Several evaluation tasks on low-resolution FER datasets were conducted and achieved competitive performance compared with previous works. The proposed model produces more efficiency than the mobile benchmark model. Additionally, it demonstrates that integration with face detection works effectively and can operate fast on low-cost devices. It also shows that the built model is feasible to be implemented in real-world scenarios. The development of classifiers can improve the model's performance in future work.

ACKNOWLEDGEMENTS

This research is funded by Fundamental Research of UNSRAT Excellence Cluster 1 (RDUU_K1) Sam Ratulangi University 2023 with contract number 334/UN12.13/LT/2023.

REFERENCES

[1] X. Liu, X. Cheng, and K. Lee, "GA-SVM-Based Facial Emotion Recognition Using Facial Geometric Features," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11532–11542, 2021, doi: 10.1109/JSEN.2020.3028075.

[2] B. Hdioud and M. E. H. Tirari, "Facial expression recognition of masked faces using deep learning," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 921–930, 2023, doi: 10.11591/ijai.v12.i2.pp921-930.

[3] S. Subudhiray, H. K. Palo, and N. Das, "K-nearest neighbor based facial emotion recognition using effective features," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 1, pp. 57–65, 2023, doi: 10.11591/ijai.v12.i1.pp57-65.

[4] 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, doi: 10.1109/TII.2022.3145862.

[5] M. D. Putro, L. Kurnianggoro, and K.-H. Jo, "High Performance and Efficient Real-Time Face Detector on Central Processing Unit Based on Convolutional Neural Network," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 4449–4457, 2021, doi: 10.1109/TII.2020.3022501.

[6] M. R. S. S. Devi, V. R. V. Kumar, and P. Sivakumar, "A Review of image Classification and Object Detection on Machine learning and Deep Learning Techniques," in *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2021, pp. 1–8, doi: 10.1109/ICECA52323.2021.9676141.

[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, doi: 10.1109/CVPR.2016.90.

[8] A. Kherraki, M. Maqbool, and R. El Ouazzani, "Efficient lightweight residual network for real-time road semantic segmentation," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 1, pp. 394–401, 2023, doi: 10.11591/ijai.v12.i1.pp394-401.

[9] Y. Nan, J. Ju, Q. Hua, H. Zhang, and B. Wang, "A-MobileNet: An approach of facial expression recognition," *Alexandria Engineering Journal*, vol. 61, no. 6, pp. 4435–4444, 2022, doi: 10.1016/j.aej.2021.09.066.

[10] F. Ma, B. Sun, and S. Li, "Facial Expression Recognition With Visual Transformers and Attentional Selective Fusion," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1236–1248, 2023, doi: 10.1109/TAFFC.2021.3122146.

[11] Z. Zhao, Q. Liu, and S. Wang, "Learning Deep Global Multi-Scale and Local Attention Features for Facial Expression Recognition in the Wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 6544–6556, 2021, doi: 10.1109/TIP.2021.3093397.

[12] 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, doi: 10.1109/TPAMI.2019.2913372.

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Sep. 2018, pp. 1–17, doi: 10.48550/arXiv.1807.06521.

[14] J. Fu et al., "Dual Attention Network for Scene Segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3141–3149, doi: 10.1109/CVPR.2019.00326.

[15] L. Lo, B. K. Ruan, H. H. Shuai, and W. H. Cheng, "Modeling Uncertainty for Low-Resolution Facial Expression Recognition," *IEEE Transactions on Affective Computing*, pp. 1–12, 2023, doi: 10.1109/TAFFC.2023.3264719.

[16] C. Shan, S. Gong, and P. W. McOwan, "Recognizing facial expressions at low resolution," in *IEEE Conference on Advanced Video and Signal Based Surveillance*, 2005, pp. 330–335, doi: 10.1109/AVSS.2005.1577290.

[17] P. N. R. Bodavarapu and P. V. V. S. Srinivas, "Facial expression recognition for low resolution images using convolutional neural networks and denoising techniques," *Indian Journal of Science and Technology*, vol. 14, no. 12, pp. 971–983, 2021, doi: 10.17485/ijst/v14i12.14.

[18] F. Nan et al., "Feature super-resolution based Facial Expression Recognition for multi-scale low-resolution images," *Knowledge-Based Systems*, vol. 236, 2022, doi: 10.1016/j.knosys.2021.107678.

[19] Y. Yan, Z. Zhang, S. Chen, and H. Wang, "Low-resolution facial expression recognition: A filter learning perspective," *Signal Processing*, vol. 169, 2020, doi: 10.1016/j.sigpro.2019.107370.

[20] Y. Song, B. He, and P. Liu, "Real-Time Object Detection for AUVs Using Self-Cascaded Convolutional Neural Networks," *IEEE Journal of Oceanic Engineering*, vol. 46, no. 1, pp. 56–67, 2021, doi: 10.1109/JOE.2019.2950974.

[21] C. S. Won, "Multi-Scale CNN for Fine-Grained Image Recognition," *IEEE Access*, vol. 8, pp. 116663–116674, 2020, doi: 10.1109/ACCESS.2020.3005150.

[22] P. Ramachandran, B. Zoph, and Q. V Le, "Searching for activation functions," in *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings*, 2018, pp. 1–13, doi: 10.48550/arXiv.1710.05941.

[23] M. G. Calvo and D. Lundqvist, "Facial expressions of emotion (KDEF): Identification under different display-duration conditions," in *Behavior Research Methods*, 2008, vol. 40, no. 1, pp. 109–115, doi: 10.3758/BRM.40.1.109.

[24] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, 2019, doi: 10.1109/TIP.2018.2868382.

[25] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 279–283, doi: 10.1145/2993148.2993165.

[26] M. Gochoo, T. H. Tan, F. Alnajjar, J. W. Hsieh, and P. Y. Chen, "Lownet: Privacy Preserved Ultra-Low Resolution Posture Image Classification," in *Proceedings - International Conference on Image Processing, ICIP*, 2020, pp. 663–667, doi: 10.1109/ICIP40778.2020.9190922.

[27] X. Zhu, Z. Li, X. Li, S. Li, and F. Dai, "Attention-aware perceptual enhancement nets for low-resolution image classification," *Information Sciences*, vol. 515, pp. 233–247, Apr. 2020, doi: 10.1016/j.ins.2019.12.013.

[28] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing Uncertainties for Large-Scale Facial Expression Recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6896–6905, doi: 10.1109/CVPR42600.2020.00693.

[29] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into Ambiguity: Latent Distribution Mining and Pairwise Uncertainty Estimation for Facial Expression Recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6244–6253, doi: 10.1109/CVPR46437.2021.00618.

[30] Y. Zhang, C. Wang, and W. Deng, "Relative Uncertainty Learning for Facial Expression Recognition," in *Advances in Neural Information Processing Systems*, 2021, vol. 21, pp. 17616–17627.

BIOGRAPHIES OF AUTHORS

Muhamad Dwisnanto Putro received a bachelor's of engineering (S.T.) in electrical engineering from Sam Ratulangi University in Manado, Indonesia, in 2010. He received an M.Eng. degree from the Department of Electrical Engineering at Gadjah Mada University in Yogyakarta, Indonesia, in 2012. He graduated a Ph.D. degree with the Department of Electrical, Electronic, and Computer Engineering, University of Ulsan, South Korea, in 2022. In 2013, he joined the Department of Electrical Engineering, Sam Ratulangi University, as an Assistant Professor. His current research interests include computer vision and deep learning, which focuses on robotic vision and perception.He can be contacted at email: dwisnantoputro@unsrat.ac.id.

Jane Litouw received a bachelor's of engineering (S.T.) in electrical engineering from Sam Ratulangi University in Manado, Indonesia, in 2003. He received an Magister Teknik (M.T) degree from STEI ITB in Bandung, Indonesia, in 2014. In 2005 she joined the Department of Electrical Engineering, Sam Ratulangi University, as lecturer. Her current research interests are fuzzy logic system, image processing and deep learning. She can be contacted by email: jane_litouw@unsrat.ac.id.

Vecky Canisius Poekoel received a bachelor's of engineering (S.T.) in electrical engineering from Institut Teknologi Sepuluh November (ITS) in Surabaya, Indonesia, in 1994. He received an M.T. degree from the Department of Electrical Engineering at Institut Teknologi Bandung (ITB), in Bandung, Indonesia, in 2005. He graduated a Dr.Eng with the Department of Computer Science and Electrical Engineering, Kumamoto University, Kumamoto, Japan, in 2014. In 1994, he joined the Department of Electrical Engineering, Sam Ratulangi University. His current research interests include Control Engineering and Artificial intelligence.He can be contacted at email: vecky.poekoel@unsrat.ac.id.