IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 July 2022, accepted 12 August 2022, date of publication 22 August 2022, date of current version 31 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3200762

RESEARCH ARTICLE

Real-Time Implementation of Face Recognition and Emotion Recognition in a Humanoid Robot Using a Convolutional Neural Network

SUCI DWIJAYANTI, (Member, IEEE), MUHAMMAD IQBAL, AND BHAKTI YUDHO SUPRAPTO, (Member, IEEE)
Department of Electrical Engineering, Universitas Sriwijaya, Indralaya 30662, Indonesia

Corresponding author: Suci Dwijayanti (sucidwijayanti@ft.unsri.ac.id)

This work was supported by DIPA of Public Service Agency, Universitas Sriwijaya, in 2022 (In accordance with the rector's decree number: 0109/UN9.3.1/SK/2022, in April 2022) in December 2021, under Grant SP DIPA-023.17.2.677515/2022.

ABSTRACT Robots can mimic humans, including recognizing faces and emotions. However, relevant studies have not been implemented in real-time humanoid robot systems. In addition, face and emotion recognition have been considered separate problems. This study proposes a combination of face and emotion recognition for real-time application in a humanoid robot. Specifically, face and emotion recognition systems are developed simultaneously using convolutional neural network architectures. The model is compared to well-known architectures, such as AlexNet and VGG16, to determine which is better for implementation in humanoid robots. Data used for face recognition are primary data taken from 30 electrical engineering students after preprocessing, resulting in 18,900 data points. Emotion data of surprise, anger, neutral, smile, and sad are taken from the same respondents and combined with secondary data for a total of 5,000 data points for training and testing. The test is carried out in real time on a humanoid robot using the two architectures. The face and emotion recognition accuracy is 85% and 64%, respectively, using the AlexNet model. VGG16 yields recognition accuracies of 100% and 73%, respectively. The proposed model architecture shows 87% and 67% accuracies for face recognition and emotion recognition, respectively. Thus, VGG16 performs better in recognizing faces as well as emotions, and it can be implemented in humanoid robots. This study also provides a method for measuring the distance between the recognized object and robot with an average error rate of 2.52%.

INDEX TERMS Accuracy, convolutional neural network, emotion recognition, face recognition, humanoid robot.

I. INTRODUCTION

Recently, the rapid growth of technology has advanced research in the field of robotics, including research on humanoid robots. A humanoid robot is a human-shaped robot equipped with a body, hands, head, and so on. Usually, a humanoid robot has the capability to interact with humans, such as recognizing the human and responding to commands given by the human.

Humanoid robots are usually expected to be socially assistive robots. Thus, face recognition is an important matter in human–machine interaction. Robots capture a human's face through a camera embedded as the eyes.

Face recognition is a technology with the capability of identifying or verifying the subject's identity in the form of images or video [1]. Some technologies have been developed

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

to recognize faces. A.E. Omer and A. Khuran implemented facial recognition using principal component analysis (PCA) [2]. Then, Ebeid [3] compared two methods, multilayer perceptron and a radial-based function with eigenface feature extraction, for face recognition. Sanjaya *et al.* [4] developed a social robot that can recognize and track the human face. In their study, they used the cascade classification method (Viola–Jones method) and a local binary pattern histogram (LBPH). However, only a few samples were used, and illumination conditions were not considered. Cilmi and Mercimek [5] also used the Haar cascade classifier to detect faces and the Kanade–Lucas–Tomasi feature tracker by considering neck movement. Zhao and Wei [6] proposed an LBPH algorithm based on neighborhood gray median (MLBPH) to improve LBPH in terms of illumination, expression, and attitude deflection. In another study, Zhi and Liu [7] used PCA to extract the features in grayscale images, a genetic algorithm to optimize the network weights of face features, and a support vector machine (SVM) as a classifier [8]. Borkar and Kuwelkar [8] also utilized PCA in combination with linear discriminant analysis (LDA) to reduce dimensionality. This method was implemented on the AT&T database. Fontaine *et al.* [9] modified the robust sparse coding algorithm to recognize labeled faces in the Wild database.

These methods depend on the face data features. Thus, feature extraction is crucial in determining the success of recognition. However, feature extraction may reduce the dimensionality of the original dataset, which may remove some important information. Thus, in recent studies, deep learning algorithms, such as convolutional neural networks (CNNs) [10] and deep CNNs [11], have been used to improve face recognition accuracy.

In addition to face recognition, emotion recognition can also be considered a major ability of machines in human–machine communication [12], [13]. Emotion recognition can be performed based on speech and facial expressions [14] as well as text [15]. For human–robot interaction, facial expressions are very important since they can carry various pieces of information [16]. Nicolai and Choi [17] discussed facial emotion recognition in the context of a fuzzy system. Adeyanju *et al.* [18] evaluated the performance of different support vector engine kernels for facial emotion recognition. Ahmed *et al.* [19] performed facial emotion recognition methods using a CNN and data augmentation by combining various datasets. Ruiz-Garcia *et al.* [20] combined a CNN and SVM to recognize emotions using the KDFF dataset. In their study, Faria *et al.* [21] used a geometrical feature based on log-covariance and angles formed by facial landmarks. Other studies considered static images [22], [23] or implemented deep learning to support emotion recognition in humanoid robots. For example, Mehendale [24] used a CNN, Li *et al.* [16] utilized a combination of a CNN and long short-term memory (LSTM), and [25] proposed a conditional generative adversarial network. [26] proposed a method of recognizing the face expression using a deep convolutional neural networks model which has input coming from the local gravitational force descriptor as features. Meanwhile, [27] used a convolutional neural network to distinguish facial expressions, namely FER-net which has been tested in five datasets.

These methods perform quite well in face recognition and facial emotion recognition tasks but are still limited since face recognition and emotion recognition have not been combined into a single recognition system, and were instead considered as different cases. However, face recognition and emotion recognition should be implemented as a unit to improve the capacity for human–robot interaction. In addition, the importance of the position of a human as an object to interact with was not considered in these studies. Furthermore, only a few studies have implemented face recognition or emotion recognition in real time [5], [16], [21], [28], [29]. Thus, this study aims to treat face recognition and emotion recognition as a unit so that robots can interact with humans by recognizing their names and emotions in real time as well as their position. Different from other studies that utilized a previously collected dataset, this study used primary data obtained from male and female students, where some students wore glasses and some female students wore a hijab. The contributions of this study are as follows:

a. Face recognition and emotion recognition are combined into one unit, and the recognition system is embedded in the robot so that it can interact with a human based on his or her face and emotions in real time.
b. The performances of well-known CNN architectures, i.e., VGG16 and AlexNet, are compared with that of the proposed modified architecture.
c. A method is proposed for measuring the distance between the object's face and the position of the robot so that the robot can determine where the human is.

This paper is structured as follows. Section 2 provides the method used in this study. The results and discussion are presented in Section 3. Finally, this paper is concluded in Section 4.

II. METHODS

A. HARDWARE DESIGN

In this study, several pieces of hardware are used to support the implementation of a humanoid robot:

1. Webcam
2. JX Servo 60KG
3. Arduino
4. Raspberry Pi
5. Dot matrix

The positions of the components used in this study are shown in Figure 1 (A-D).

The face images are captured using a webcam embedded as the robot's eyes. A dot matrix is used to present characters such as lines and circles. Such characters represent the form of eyes. JX Servo functions as the neck of the robot so it

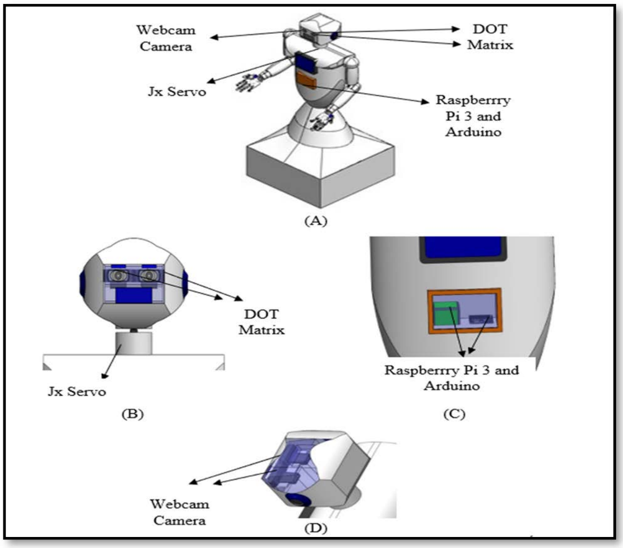


FIGURE 1. (A-D). Component design on a humanoid robot.

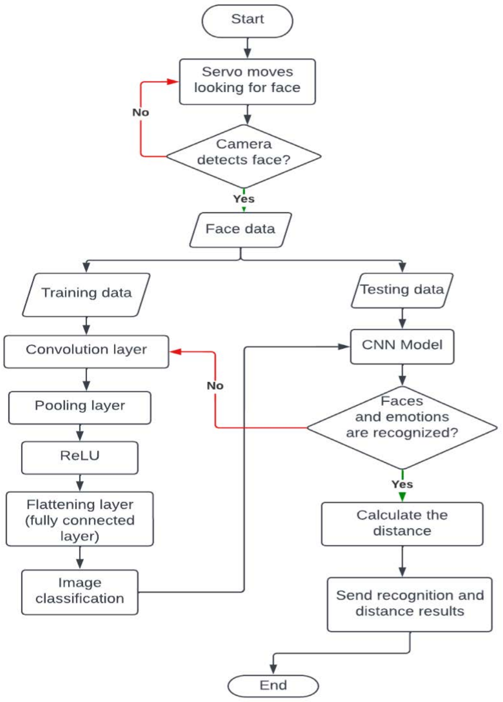


FIGURE 2. Design of the recognition system.

can move to follow the position of the human's face after detecting and recognizing it.

B. FACE AND EMOTION RECOGNITION ALGORITHM SYSTEM DESIGN

The design of the recognition system embedded in the robot for recognizing a person's face is shown in Figure 2.

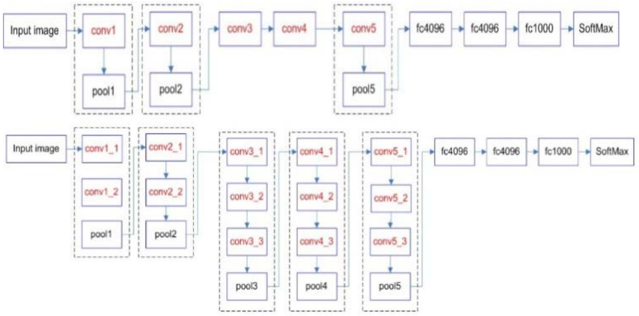


FIGURE 3. AlexNet architecture (above) and VGG16 (below) [34].

TABLE 1. Parameters of the proposed modified architecture.

Parameter	*Value*
Dropout after pooling layer	0.05, 0.1
Dropout fully connected layer	0.25, 0.1
Dense layer	64, 128
Learning rate	0.0001
Batch size	16

As shown in Figure 2, the first stage is camera detection of the face to obtain the face images as a dataset for training. The CNN used in this study consists of AlexNet [30] and VGG16 [31]. These architectures are chosen because they have shown good performance in face recognition [32] and emotion recognition [33]. The architecture of AlexNet and VGG16 can be seen in Figure 3.

In addition, the AlexNet architecture is modified by changing some parameters, as shown in Table 1.

In addition to face and emotion recognition, the distance is also measured to detect the position of the object. Specifically, the x and y coordinates between faces/emotions and the humanoid robot are measured. When recognizing faces and emotions, the frame represents 4 variables (x, y, w, and h), where x and y are the bounding boxes x for the upper-left side and y for the lower-right side and w and h are the width and height of the bounding box, which are processed to obtain the end x and end y values to determine the values of kdX and kdY. The calculation of coordinates x and y is as follows:

$$kdX = \frac{StartX + EndX}{2} \tag{1}$$

$$kdY = \frac{StartY + EndY}{2} \tag{2}$$

where kdX is the x coordinate and kdY is the y coordinate, StartX is the starting point of the X-axis in the bounding box, StartY is the starting point of the Y-axis in the bounding box, EndX is the end point of the X-axis in the bounding box, and EndY is the end point of the Y-axis in the bounding box.

Additionally, the distance from the recognized face to the camera embedded in the robot is calculated as follows:

$$focal\ length = \frac{w \times d}{W} \tag{3}$$

$$distance = \frac{W \times f}{w} \tag{4}$$

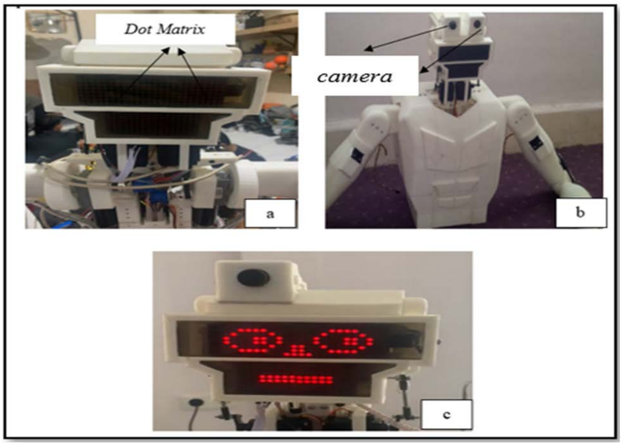


FIGURE 4. (a-c). Humanoid robot.

where f is the focal length, w is the width in pixels, d is the distance in cm, and W is the width in cm.

The values of the x and y coordinates as well as distance are essential when the humanoid robot interacts with the human.

C. SYSTEM EVALUATION

Evaluation is performed to determine the performance of the developed face and expression recognition system. Performance includes accuracy measured as the recognition rate of the proposed system for faces and emotions in real time. The formula for calculating the accuracy of the test is presented in formula (1):

$$\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN}, \quad (5)$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and TN represents true negatives. This formula is used to obtain face recognition or emotion recognition accuracy. A calculation of the accuracy value shows the level of effectiveness per class of a classification.

III. RESULTS AND DISCUSSION

A. IMPLEMENTATION OF THE HUMANOID ROBOT

The humanoid robot includes mechanical design and overall wiring of the components used, such as the dot matrix, which is used to display the visual appearance of the eyes of the humanoid robot, and the JX Servo, which is used to move the head of the humanoid robot when looking for the facial position of a person to be recognized, and a camera module connected to a laptop that is used to perform face and emotion recognition. The specifics of the humanoid robot design are shown in Figure 4(a-c).

B. DATASET COLLECTION

In this study, facial datasets were obtained from 30 Universitas Sriwijaya students, consisting of 21 male students and 9 female students. All the students gave their permission to use their face data. An example of face data is provided in Figure 5. The data were obtained using a webcam with a resolution of 640 × 480 pixels. The data were then processed and extracted to perform face recognition and emotion recognition on the images. The primary dataset includes 50 data points per class. Furthermore, the data were processed and extracted to produce 18,900 data points, with each class totaling 630 data points. The percentage of data used as training data is 80%, i.e., 15,120 data points, and that of the test data is 20%, i.e., 3,780 data points.

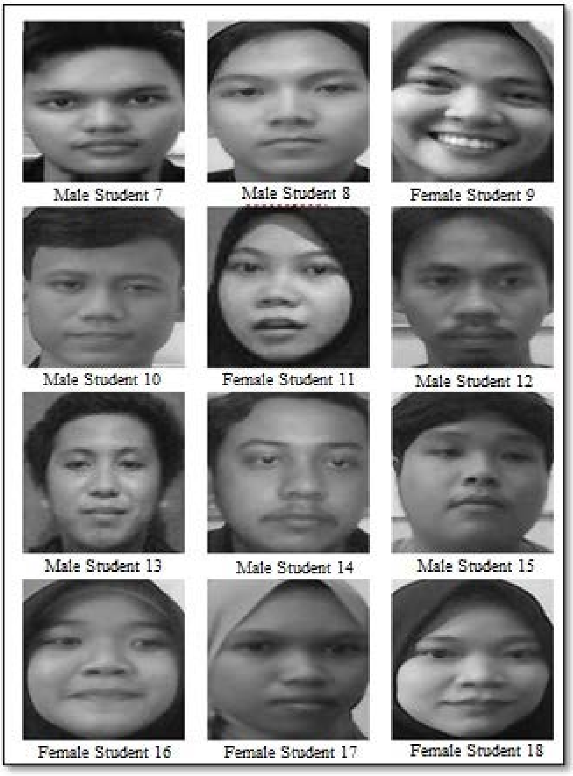


FIGURE 5. Some face samples.

Emotion data were also taken from the same respondents. To add variation to the data, the dataset from Kaggle [35] is used in this study. Examples of emotions from Kaggle can be seen in Figure 6. The facial emotions used in this study consist of 5 expressions, namely, smile, anger, surprise, neutral and sad. A total of 4,000 training data points are used, and 1,000 test data points were used.

The collected training data and test data are then preprocessed. This preprocessing includes cropping each image to remove any background or regions other than the face.

After the cropping process, image resizing is carried out by changing the image size from 640 × 480 pixels to 120 × 120 pixels. This is done to reduce the size of the dataset used during training. Then, the next process is to perform image augmentation to make the image variations fit a range = 0.1, shear_range = 0.9, width shift range = 1 and height shift range = 0.01.

TABLE 2. Face recognition.

No.	Name	Model A (AlexNet) 500 epochs	Model B (VGG16) 500 epochs	Model C (Proposed model) 500 epochs
1.	MS 1	Recognized	Recognized	Recognized
2.	MS 2	Unrecognized	Recognized	Recognized
3.	MS 3	Recognized	Recognized	Recognized
4.	MS 4	Recognized	Recognized	Recognized
5.	FS 5	Recognized	Recognized	Recognized
6.	MS 6	Recognized	Recognized	Recognized
7.	MS 7	Recognized	Recognized	Recognized
8.	MS 8	Unrecognized	Recognized	Recognized
9.	FS 9	Recognized	Recognized	Recognized
10.	MS 10	Unrecognized	Recognized	Recognized
11.	FS 11	Recognized	Recognized	Recognized
12.	MS 12	Recognized	Recognized	Recognized
13.	MS 13	Recognized	Recognized	Recognized
14.	MS 14	Recognized	Recognized	Recognized
15.	MS 15	Recognized	Recognized	Recognized
16.	FS 16	Recognized	Recognized	Recognized
17.	FS 17	Recognized	Recognized	Recognized
18.	FS 18	Unrecognized	Recognized	Recognized
19.	MS 19	Recognized	Recognized	Recognized
20.	MS 20	Recognized	Recognized	Recognized
21.	MS 21	Recognized	Recognized	Recognized
22.	FS 22	Recognized	Recognized	Recognized
23.	FS 23	Recognized	Recognized	Recognized
24.	MS 24	Recognized	Recognized	Recognized
25.	MS 25	Recognized	Recognized	Recognized
26.	MS 26	Recognized	Recognized	Recognized
27.	MS 27	Recognized	Recognized	Recognized
28.	FS 28	Recognized	Recognized	Recognized
29.	MS 29	Recognized	Recognized	Recognized
30.	MS 30	Recognized	Recognized	Recognized
Accuracy		86%	100%	95%

Note: MS, male student; FS, female student

TABLE 3. Emotion recognition.

No.	Emotion	Model A (AlexNet) 500 *epochs*	Model B (VGG16) 500 *epochs*	Model C (Proposed model) 500 *epochs*
1.	Surprise	Recognized	Recognized	Recognized
2.	Angry	Unrecognized	Recognized	Recognized
3.	Neutral	Recognized	Recognized	Recognized
4.	Sad	Recognized	Recognized	Recognized
5.	Smile	Recognized	Recognized	Recognized
Accuracy		64%	82%	71%

C. FACE RECOGNITION

In this study, the proposed architecture model (model C) was compared with AlexNet (model A) and VGG16 (model B) using 500 epochs. The training loss for each class can be seen in Figure 7. As shown in Figure 7, the VGG16 architecture has a lower training loss than AlexNet and the proposed model alone. The training losses of the proposed architecture and AlexNet are 0.011, 0.151 and 0.052, respectively. These results indicate that the VGG16 architecture can provide better performance than AlexNet and its model for face recognition. Nevertheless, the average loss for model C is close to model B, which is 0,052. This indicates that the proposed model architecture (model C), which is a modification of AlexNet, has a smaller loss compared to AlexNet. In addition, model C has smaller parameters, so its training time is faster than that of model B.

Table 2 shows the accuracy obtained with the test data. The model using the VGG16 architecture performed much better than AlexNet and our model. The VGG16 architecture recognizes all test data samples, and our model can recognize the face with an accuracy of 95%. These results indicate that the VGG16 model and our model can recognize faces well. Additionally, AlexNet recognizes 86% of the test data. An error occurred when recognizing samples 2, 8, 10, and 18. The cause of the error is the similarity of the facial features. For example, male student 8 has a similar face to male student 29, as shown in Figure 8.

TABLE 4. Face and emotion recognition on 10 data samples in real time with the VGG16 model.

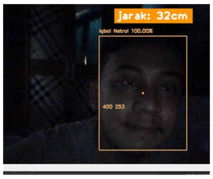
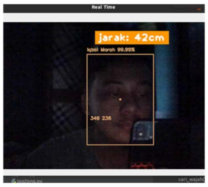
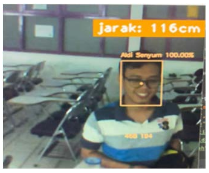
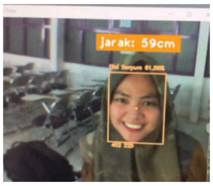
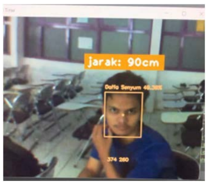
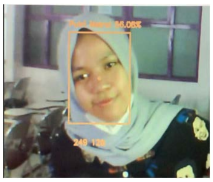
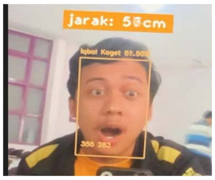
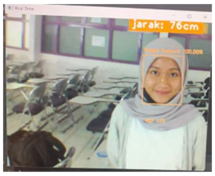
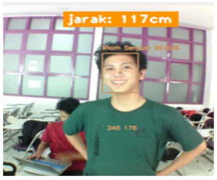
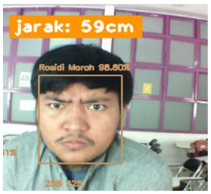
No.	Name & Emotion	Image Detection and Recognition	Lighting Condition	Distance	Real Distance	Distance Difference	Description	
							Face	Emotion
1.	MS14 Neutral		Dark	32 cm	30 cm	2 cm	Recognized	Recognized
2.	MS14 Neutral	 Real Time jarak: 42cm Iqbal Marah 99.99% 349 236 cari_wajah(	Dark	42 cm	40 cm	2 cm	Recognized	Unrecognized
3.	MS3 Smile		Dim	116 cm	115 cm	1 cm	Recognized	Recognized
4.	FS9 Smile		Dim	59 cm	60 cm	1 cm	Recognized	Recognized
5.	MS7 Angry		Dim	90 cm	90 cm	-	Recognized	Unrecognized
6.	FS22 Smile		Dim	46 cm	45 cm	1 cm	Recognized	Unrecognized
7.	MS14 Surprise		Bright	57 cm	60 cm	3 cm	Recognized	Recognized
8.	FS28 Smile		Bright	76 cm	75 cm	1 cm	Recognized	Recognized
9.	MS13 Smile		Bright	117 cm	115 cm	2 cm	Recognized	Recognized

TABLE 4. ***(Continued.)*** **Face and emotion recognition on 10 data samples in real time with the VGG16 model.**

10.	MS26 Angry		Bright	59 cm	60 cm	1 cm	Recognized	Recognized

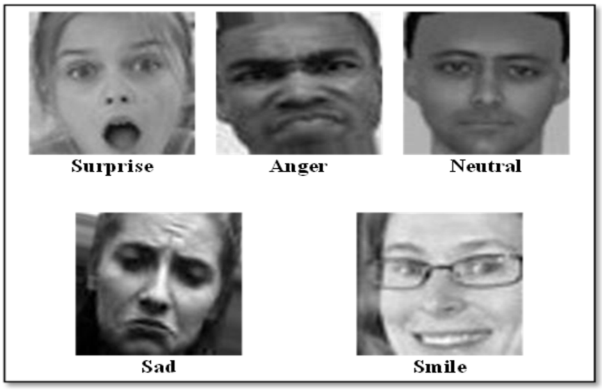


FIGURE 6. **5 Sample of emotions (surprise, anger, neutral, sad, and smile).**

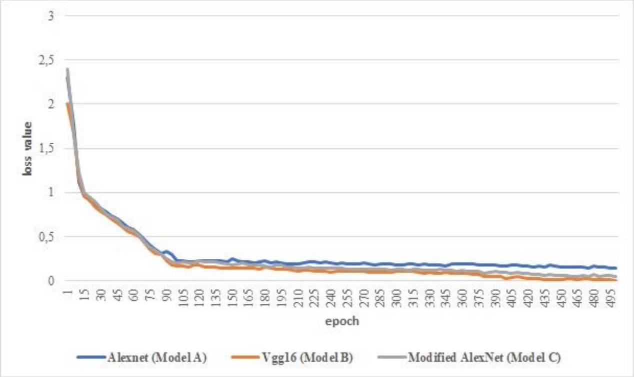


FIGURE 7. **Training loss in face recognition.**

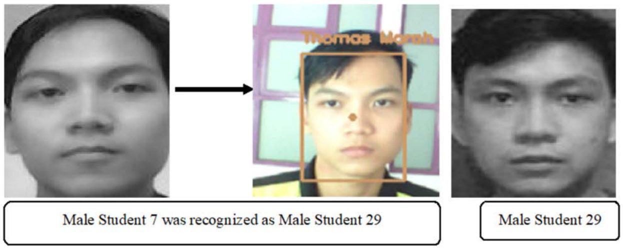


FIGURE 8. **Examples of similar faces.**

D. EMOTION RECOGNITION

The training loss results for emotion recognition are shown in Figure 9. The training losses for AlexNet, VGG16 and our model architecture were 0.352, 0.1875 and 0.301, respectively. These results show that the VGG16 architecture is superior to AlexNet and our model. This may be due to the parameters used in the proposed architecture and the number of layers contained in each architecture.

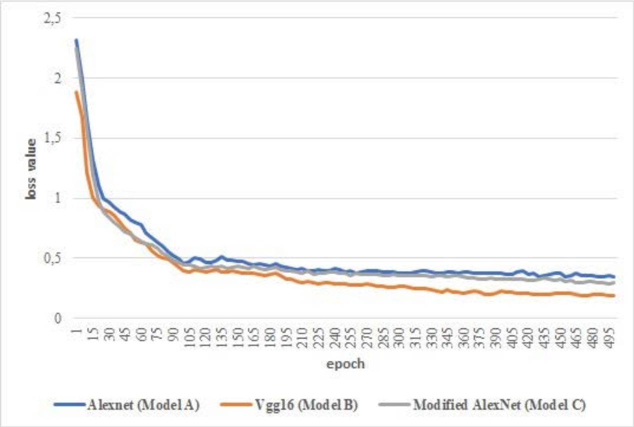


FIGURE 9. **Training loss in emotion recognition.**

Table 3 shows the accuracy obtained from the test data. As shown in the table, the emotion recognition accuracy is lower than that of face recognition. This may be caused by the similarity of the emotions of each individual. As shown in Figure 9, the emotion of anger is similar to surprise. The best accuracy is obtained by model B (VGG16), with an accuracy of 82%. Additionally, the accuracy of the proposed model (model C) is 71%. It is better than model A (AlexNet) and lower than model B. Even though its accuracy is lower than model B, the proposed model has a smaller number of layers, so the training process is faster compared to model B. These results indicate that the VGG16 model and our model can recognize emotions well.

E. REAL-TIME TESTING ON THE HUMANOID ROBOT

Real-time testing was carried out directly on the camera module attached to the humanoid robot. This test is conducted in a room with three conditions of illumination: dark, dim, and bright. This test aimed to determine whether the humanoid robot can detect or recognize the faces and emotions of someone around it with input in the form of face and emotion detection results from the VGG16 model (model B) and

TABLE 5. FACE and emotion recognition on 10 data samples in real time with the proposed model.

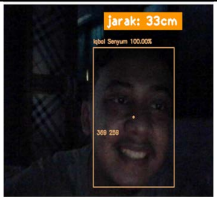
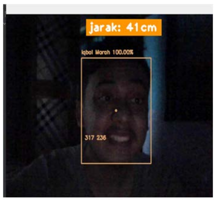
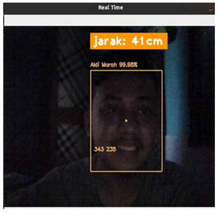
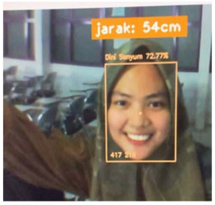
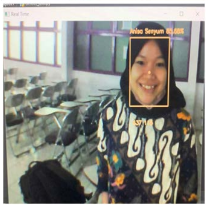
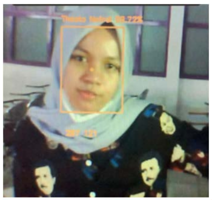
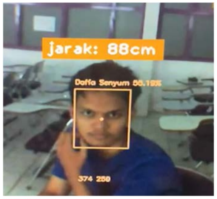
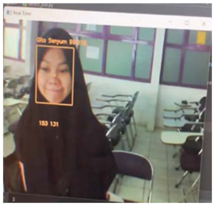
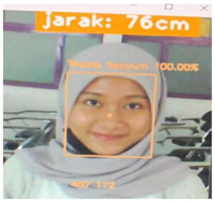
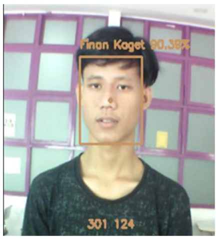
No.	Name & Emotion	Image Detection and Recognition	Lighting Condition	Distance	Real Distance	Distance Difference	Description	
							Face	Emotion
1.	MS14 Neutral		Dark	33 cm	30 cm	3 cm	Recognized	Recognized
2.	MS14 Smile		Dark	41 cm	40 cm	1 cm	Recognized	Unrecognized
3.	MS14 Smile		Dark	41 cm	40 cm	1 cm	Unrecognized	Unrecognized
4.	FS9 Smile		Dim	54 cm	55 cm	1 cm	Recognized	Recognized
5.	FS18 Smile		Dim	50 cm	50 cm	-	Unrecognized	Recognized
6.	FS22 Neutral		Dim	46 cm	45 cm	1 cm	Unrecognized	Recognized
7.	MS7 Angry		Dim	88 cm	90 cm	2 cm	Recognized	Unrecognized
8.	FS11 Neutral		Dim	56 cm	55 cm	1 cm	Recognized	Unrecognized

TABLE 5. ***(Continued.)*** **FACE and emotion recognition on 10 data samples in real time with the proposed model.**

9.	FS28 Smile		Bright	76 cm	75 cm	1 cm	Recognized	Recognized
10.	MS10 Neutral		Bright	59 cm	60 cm	1 cm	Recognized	Unrecognize d

the proposed model (model C), which are stored during the training process with 500 training epochs.

The recognized data, coordinates, and distance are sent using a robot operating system (ROS). Such data is needed later for the voice recognition and movement systems of the robot.

The system testing results using the VGG16 model (model B) with 500 epochs can be seen in Table 4.

In tests 1 to 10, the movement of the humanoid robot in looking for faces and then detecting and recognizing faces is still not stable because the camera position must continuously move to follow the face position of the person to be recognized. However, the humanoid robot can still detect and recognize the person in front of it. Face recognition using model B is good because the system can recognize the faces of people in front of it, but there are still errors in recognizing emotions. In the 2nd, 5th and 9th tests, there are still system errors in detecting and recognizing emotions. This is due to several factors, such as poor or dim lighting conditions, the distance between the robot and the person to be recognized being quite far, the face not being positioned toward the camera and the lack of variation in training data from various possible conditions that exist during real-time testing on the humanoid robot. Additionally, in model B, the training process time is quite long due to the large number of layers in model B. The percentage of successes in recognizing faces using model B for all 30 students was 100% and that for emotion recognition was 73%.

The table also shows that the study can measure the distance between the recognized object and the position of the robot well, with an average error rate of 2.52%. This error may be caused by the changing position of the object, causing the distance measurement accuracy to decrease.

The results of system testing using the proposed model (model C) with 500 epochs can be seen in Table 5.

Similar to testing using the VGG16 model (model B), in tests 1 to 11, the movement of the humanoid robot in finding faces and then detecting and recognizing faces is still not stable because the camera position must continuously move to follow the position of the face of the person to be recognized. Even so, the humanoid robot can still detect and recognize the person in front of it. In the 2nd, 3rd, 7th, 8th, and 10th tests, there are still system errors in detecting and recognizing emotions, and in the 3rd, 5th and 6th tests, there are still system errors in detecting and recognizing faces. This can be due to several factors, such as poor or dim lighting conditions, the distance between the robot and the person to be recognized being quite far, and a lack of variation in training data from various possible conditions that exist during real-time testing on humanoid robots. Overall, the success percentages for recognizing 30 students' faces and emotions using model C were 87% and 67%, respectively.

The test results obtained using model B and model C in real terms show that the implementation of face and emotion recognition in a CNN-based humanoid robot with model B and model C architectures is successfully carried out using system input in the form of point coordinates of the detected and recognized face to move the servo and dot matrix with 100% accuracy for facial recognition and 73% for emotion recognition in model B. Additionally, the accuracy obtained by the proposed model (model C) is 87% and 67% for face and emotion recognition, respectively. Although model B performs better in this study, the training process time is much longer than that for model C due to the number of layers used. However, the movement of the servo can also affect the accuracy of real-time recognition, and lighting is also very important during real-time recognition.

The real-time experiments shown in Tables 4 and 5 also show that this study can calculate the distance between the robot and object well. The distance and illumination may influence the accuracy of recognizing faces and emotions. The VGG16 model and the proposed model are quite good for recognizing faces in bright, dim or dark rooms. However, the emotion is rather difficult to recognize, especially for the dark and dim room. Additionally, a farther distance between the object and the robot may cause difficulty in recognizing faces and emotions.

IV. CONCLUSION

Based on this research, it can be concluded that the VGG16 model (model B) is superior to model C and model A, as shown by the success rates in detecting and recognizing faces and emotions of 100% and 73%, respectively. The smallest average loss of face and emotion recognition for VGG16 is 0.011 and 0.1875, respectively. However, the training process carried out by model B is much longer than that of model A and model C because the number of layers used is much larger than that in models A and C. The face and emotion recognition results obtained with model C are not much different from those obtained with model B and model C. Model C has the advantage of a faster training process because of fewer parameters.

Models B and C were then used as inputs to the humanoid robot's face and emotion recognition system. This study revealed that the implementation and development of a CNN with the VGG16 architecture (model B) and a modified AlexNet model (model C) for recognizing faces and emotions as input to a humanoid robot was successful. This was shown by the accuracy values obtained, with success percentages of 100% and 73% for face and emotion recognition, respectively, when applying the VGG16 architecture. In model C, the corresponding accuracies were 87% and 67%. In addition, the face and emotion recognition processes carried out in this study were combined into one recognition framework.

This study also showed that recognition can be implemented in real time for humanoid robots and that the distance between the object and the humanoid robot can be measured well. The distance and illumination are also important factors in recognition. Thus, it is necessary to upgrade the size of the training dataset while considering illumination in future studies. In addition, the proposed system still needs to be repaired and upgraded in future work.

REFERENCES

[1] D. S. Trigueros, L. Meng, and M. Hartnett, "Face recognition: From traditional to deep learning methods," 2018, *arXiv:1811.00116*.

[2] A. E. Omer and A. Khurran, "Facial recognition using principal component analysis based dimensionality reduction," in *Proc. Int. Conf. Comput., Control, Netw., Electron. Embedded Syst. Eng. (ICCNEEE)*, Sep. 2015, pp. 434–439.

[3] H. M. Ebeid, "Using MLP and RBF neural networks for face recognition: An insightful comparative case study," in *Proc. Int. Conf. Comput. Eng. Syst.*, Nov. 2011, pp. 123–128.

[4] W. S. M. Sanjaya, D. Anggraeni, K. Zakaria, A. Juwardi, and M. Munawwaroh, "The design of face recognition and tracking for human-robot interaction," in *Proc. 2nd Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, Nov. 2017, pp. 315–320.

[5] Y. Gizlenmistir, "Design and implementation of real time face tracking humanoid robot," in *Proc. 26th Signal Process. Commun. Appl. Conf. (SIU)*, May 2018, pp. 1–6.

[6] X. Zhao and C. Wei, "A real-time face recognition system based on the improved LBPH algorithm," in *Proc. IEEE 2nd Int. Conf. Signal Image Process. (ICSIP)*, Aug. 2017, pp. 72–76.

[7] H. Zhi and S. Liu, "Face recognition based on genetic algorithm," *J. Vis. Commun. Image Represent.*, vol. 58, pp. 495–502, Jan. 2019, doi: 10.1016/j.jvcir.2018.12.012.

[8] N. R. Borkar and S. Kuwelkar, "Real-time implementation of face recognition system," in *Proc. Int. Conf. Comput. Methodolog. Commun. (ICCMC)*, Jul. 2017, pp. 249–255.

[9] X. Fontaine, R. Achanta, and S. Susstrunk, "Face recognition in real-world images," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 1482–1486.

[10] M. Coskun, A. Ucar, O. Yildirim, and Y. Demir, "Face recognition based on convolutional neural network," in *Proc. Int. Conf. Modern Electr. Energy Syst. (MEES)*, Nov. 2017, pp. 376–379.

[11] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A comprehensive study on center loss for deep face recognition," *Int. J. Comput. Vis.*, vol. 127, pp. 668–683, Jun. 2019, doi: 10.1007/s11263-018-01142-4.

[12] M. Egger, M. Ley, and S. Hanke, "Emotion recognition from physiological signal analysis: A review," *Electron. Notes Theor. Comput. Sci.*, vol. 343, pp. 35–55, May 2019, doi: 10.1016/j.entcs.2019.04.009.

[13] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, Jan. 2020, doi: 10.3390/s20030592.

[14] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 60–75, Jan. 2019, doi: 10.1109/TAFFC.2017.2713783.

[15] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019, doi: 10.1109/ACCESS.2019.2934529.

[16] T.-H.-S. Li, P.-H. Kuo, T.-N. Tsai, and P.-C. Luan, "CNN and LSTM based facial expression analysis model for a humanoid robot," *IEEE Access*, vol. 7, pp. 93998–94011, 2019, doi: 10.1109/ACCESS.2019.2928364.

[17] A. Nicolai and A. Choi, "Facial emotion recognition using fuzzy systems," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2015, pp. 2216–2221.

[18] I. A. Adeyanju, E. O. Omidiora, and O. F. Oyedokun, "Performance evaluation of different support vector machine kernels for face emotion recognition," in *Proc. SAI Intell. Syst. Conf. (IntelliSys)*, Nov. 2015, pp. 804–806.

[19] T. U. Ahmed, S. Hossain, M. S. Hossain, R. ul Islam, and K. Andersson, "Facial expression recognition using convolutional neural network with data augmentation," in *Proc. Joint 8th Int. Conf. Informat., Electron. Vis. (ICIEV) 3rd Int. Conf. Imag., Vis. Pattern Recognit. (icIVPR)*, May 2019, pp. 336–341.

[20] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, and V. Palade, "A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots," *Neural Comput. Appl.*, vol. 29, no. 7, pp. 359–373, 2018, doi: 10.1007/s00521-018-3358-8.

[21] D. R. Faria, M. Vieira, and F. C. C. Faria, "Towards the development of affective facial expression recognition for human-robot interaction," in *Proc. 10th Int. Conf. Pervasive Technol. Rel. Assistive Environ.*, Jun. 2017, pp. 300–304.

[22] J. Guo, Z. Lei, J. Wan, E. Avots, N. Hajarolasvadi, B. Knyazev, A. Kuharenko, J. C. S. Jacques, Jr., X. Baró, H. Demirel, S. Escalera, A. Allik, and G. Anbarjafari, "Dominant and complementary emotion recognition from still images of faces," *IEEE Access*, vol. 6, pp. 26391–26403, 2018, doi: 10.1109/ACCESS.2018.2831927.

[23] S. Gupta, "Facial emotion recognition in real-time and static images," in *Proc. 2nd Int. Conf. Inventive Syst. Control (ICISC)*, Jan. 2018, pp. 553–560.

[24] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *Social Netw. Appl. Sci.*, vol. 2, no. 3, p. 446, Feb. 2020, doi: 10.1007/s42452-020-2234-1.

[25] J. Deng, G. Pang, Z. Zhang, Z. Pang, H. Yang, and G. Yang, "cGAN based facial expression recognition for human-robot interaction," *IEEE Access*, vol. 7, pp. 9848–9859, 2019, doi: 10.1109/ACCESS.2019.2891668.

[26] K. Mohan, A. Seal, O. Krejcar, and A. Yazidi, "Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[27] K. Mohan, A. Seal, O. Krejcar, and A. Yazidi, "FER-Net: Facial expression recognition using deep neural net," *Neural Comput. Appl.*, vol. 33, no. 15, pp. 9125–9136, Aug. 2021.

[28] C. Tsiourti, A. Weiss, K. Wac, and M. Vincze, "Multimodal integration of emotional signals from voice, body, and context: Effects of (In)congruence on emotion recognition and attitudes towards robots," *Int. J. Social Robot.*, vol. 11, no. 4, pp. 555–573, Aug. 2019, doi: 10.1007/s12369-019-00524-z.

[29] A. G. Pour, A. Taheri, M. Alemi, and A. Meghdari, "Human–robot facial expression reciprocal interaction platform: Case studies on children with autism," *Int. J. Social Robot.*, vol. 10, no. 2, pp. 179–198, Apr. 2018, doi: 10.1007/s12369-017-0461-4.

[30] A. Krizhevsky, I. Sulskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, 2012, pp. 1–9.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[32] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. 26th Brit. Mach. Vis. Conf. (BMVC)*. Swansea, U.K., 2015, pp. 1–12.

[33] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, "Deep learning approaches for facial emotion recognition: A case study on FER-2013," in *Advances in Hybridization of Intelligent Methods: Models, Systems and Applications*, I. Hatzilygeroudis and V. Palade, Eds. Cham, Switzerland: Springer, 2018, pp. 1–16.

[34] W. Yu, K. Yang, Y. Bai, T. Xiao, H. Yao, and Y. Rui, "Visualizing and comparing AlexNet and VGG using deconvolutional layers," in *Proc. 33 rd Int. Conf. Mach. Learn.* New York, NY, USA, 2016, pp. 1–7.

[35] G. Sharma. (2019). *CK+48 5 Emotions*. Accessed: Feb. 5, 2022. [Online]. Available: https://www.kaggle.com/datasets/gauravsharma99/ck48-5-emotions

SUCI DWIJAYANTI (Member, IEEE) received the M.S. degree in electrical and computer engineering from Oklahoma State University, Stillwater, OK, USA, in 2013, and the Ph.D. degree from the Graduate School of Natural Science and Technology, Kanazawa University, Japan, in 2018. From 2007 to 2008, she was an Engineer with ConocoPhillips Indonesia Inc., Ltd. Since 2008, she has been with the Department of Electrical Engineering, Universitas Sriwijaya, Indonesia. Her research interests include signal processing and machine learning. She received a Fulbright Scholarship for her master's degree.

MUHAMMAD IQBAL was born in Palembang, South Sumatra, in 2000. He is currently pursuing the degree in electrical engineering with Universitas Sriwijaya. He has participated in various academic activities, such as being a basic Physics Laboratory Assistant and a member of the Electrical Engineering Student Association. He has participated in several regional and national contests, such as the Indonesian robot contest and student creativity program held by the Indonesian Ministry of Education and Culture, in 2021. His research interests include robotics and image processing.

BHAKTI YUDHO SUPRAPTO (Member, IEEE) was born in Palembang, South Sumatra, Indonesia, in February 1975. He is currently pursuing the Graduate degree in electrical engineering with Sriwijaya University, Indonesia. His master's and doctoral programs in electrical engineering at Universitas Indonesia (UI). He is also an Academic Staff Member of Electrical Engineering at Universitas Sriwijaya. His research interests include control and intelligent systems.