IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 19, 2019, accepted July 4, 2019, date of publication July 11, 2019, date of current version July 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2928364

CNN and LSTM Based Facial Expression Analysis Model for a Humanoid Robot

TZUU-HSENG S. LI[1], (Member, IEEE), PING-HUAN KUO[2], TING-NAN TSAI[1], AND PO-CHIEN LUAN[1]

[1]aiRobots Laboratory, Department of Electrical Engineering, National Cheng Kung University, Tainan 701, Taiwan
[2]Computer and Intelligent Robot Program for Bachelor Degree, National Pingtung University, Pingtung 90004, Taiwan

Corresponding author: Ping-Huan Kuo (phkuo@mail.nptu.edu.tw)

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST 106-2218-E-153-001-MY3 and Grant MOST 106-2221-E-006-009-MY3.

ABSTRACT Robots must be able to recognize human emotions to improve the human–robot interaction (HRI). This study proposes an emotion recognition system for a humanoid robot. The robot is equipped with a camera to capture users' facial images, and it uses this system to recognize users' emotions and responds appropriately. The emotion recognition system, based on a deep neural network, learns six basic emotions: happiness, anger, disgust, fear, sadness, and surprise. First, a convolutional neural network (CNN) is used to extract visual features by learning on a large number of static images. Second, a long short-term memory (LSTM) recurrent neural network is used to determine the relationship between the transformation of facial expressions in image sequences and the six basic emotions. Third, CNN and LSTM are combined to exploit their advantages in the proposed model. Finally, the performance of the emotion recognition system is improved by using transfer learning, that is, by transferring knowledge of related but different problems. The performance of the proposed system is verified through leave-one-out cross-validation and compared with that of other models. The system is applied to a humanoid robot to demonstrate its practicability for improving the HRI.

INDEX TERMS Convolutional neural network, long short-term memory, transfer learning, facial expression analysis.

I. INTRODUCTION

With developments in robots, studies are increasingly focusing on applications such as home service robots, health care robots, manufacturing robots, and humanoid robots. To make robots more convenient, efficient, and intelligent and to integrate them into the human society, the human–robot interaction (HRI) must be improved [1]–[2]. Emotion recognition can enable robots, machines, and computers to determine proper reactions during human interactions.

Emotions play an important role in human interactions as they let people articulate themselves without words. Emotions include cognitive appraisal, bodily language, action tendencies, expressions, and feelings [3]. People would not be able to get along with each other without emotions. Therefore, emotion recognition will certainly improve the HRI.

The associate editor coordinating the review of this manuscript and approving it for publication was Mu-Yen Chen.

Emotion recognition involves considerable information including facial expressions, body language, pitch and tone of voice, and semantics. Facial expressions are crucial because they convey considerable information that can be widely used in various applications in different fields. Furthermore, facial expressions can convey the same information across different cultures and countries.

It remains challenging for computers and robots to classify facial expressions under different light conditions, poses, and backgrounds and across people of different ages, genders, and ethnicities. In one study [4], the Facial Action Coding System (FACS) was proposed for quantifying human facial movement. This system is a practical solution for detecting facial movement within the field of behavioral science. Essentially, facial expressions are identified according to several muscle movements. On the basis of the movements of these facial muscles, the FACS decomposes the facial expressions into their component actions. Moreover, the decomposed component actions can be used for further applications [5].

The FACS is based on the simulation of facial muscle movement. An action unit (AU) is comprised of segments of the muscles involved in facial expression [6]. Seventeen major AUs are involved in basic facial expression, and all facial expressions are determined by the FACS through identification of these AUs.

Experts and scholars have developed various definitions of the basic human emotions at different points in time [7]. These definitions have been based on various considerations but are all relatively reasonable and have certain theoretical bases. In this paper, we employ the six basic emotions proposed by Ekman et al.: happiness, anger, disgust, fear, sadness, and surprise, as defined in [8]. These six emotions are also the most common definitions and are used in several relevant academic studies [9]–[13].

Facial expression recognition can be effectively achieved by analyzing static images [14], [15] or dynamic image sequences [16]–[18]. Static images can be used to extract the precise attributes of the geometric and appearance features of facial expressions; however, these features cannot describe emotions completely because static images lack dynamic factors related to facial expressions. Facial expressions can be considered a combination of contraction and relaxation of one or more facial muscles. One study [19] manually selected the most expressive frames in an image sequence for experiments. This afforded high accuracy; however, it is not a reasonable way to verify the feasibility of the method.

This study proposes an emotion recognition system based on a deep neural network to improve the HRI. To extract the geometric and appearance features of facial expressions, a convolutional neural network (CNN) [20]–[23] is trained to classify a large number of static images from a dataset.

Facial expressions are dynamic. Therefore, long short-term memory (LSTM) [24]–[28], an enhanced recurrent neural network (RNN), is used to capture the temporal and contextual information of facial expressions. Furthermore, transfer learning is used to improve the performance of the emotion recognition system [29].

Traditional machine learning has some problems and limitations. For example, when we do not have sufficient labeled training data for a given task, it is very difficult to train a model well and to achieve good performance. Another limitation of machine learning is that each model is trained for particular training data and a particular task. Therefore, the model has to be rebuilt for newly collected training data. Compared with traditional machine learning, transfer learning can improve the performance by transferring known knowledge learned from other related data.

This study (1) proposes a CNN and LSTM based model for facial emotion recognition; (2) uses the concept of transfer learning to improve the model performance; (3) develops a humanoid robot for an experiment; and (4) verifies the feasibility of the proposed model for facial emotion recognition.

The remainder of this paper is organized as follows. Section II describes related works. Section III introduces the proposed model that combines CNN and LSTM. Section IV presents experimental results demonstrating the feasibility and performance of the proposed model. Section V presents the discussions. Finally, Section VI presents the conclusions of this study.

II. RELATED WORKS

There are several works related to emotion recognition, facial expression recognition, deep neural network and transfer learning. And we can discuss them below.

A. FACIAL EXPRESSION RECOGNITION

Some studies have analyzed static images for facial expression recognition. However, facial expressions are produced by the contraction and relaxation of some facial muscles. Therefore, it is better to consider both dynamic and temporal factors in facial expressions, although some studies can extract geometric and appearance factors well.

One study [19] manually selected the most expressive frames in an image sequence for experiments. This afforded high accuracy; however, it is not a reasonable way to verify the feasibility of the method.

B. DEEP NEURAL NETWORK

Many deep learning methods have been used to solve some difficult tasks and improve performance, such as CNN, RNN [30], improved deep neural networks [31]–[33], and enhanced models such as LSTM.

CNN is known to have great ability to analyze images and to handle computer vision tasks such as classification, recognition, and identification. Long *et al.* [34] and Chen *et al.* [35] have performed semantic segmentation using deep CNN; Yu and Zhang [36] performed static-image-based facial expression recognition using deep CNNs; and one study investigated the effect of CNN depth on large-scale image recognition [37]. Therefore, many studies have effectively used CNN in facial expression recognition to extract features. To enhance the capabilities of HRI and robot–robot interaction, one study [38] proposed a CNN architecture that gives robots the ability to recognize emotions. The network has three convolution layers and one fully connected layer as the output, and information from speech, gestures, and facial recognition is employed as the CNN input. Although the CNN architecture has only four layers, it has an acceptable emotion recognition ability and performed well in experiments. However, the method can only be applied to still images because of the limitations of the CNN architecture; the method cannot be applied to continuous dynamic images.

RNN has been proposed to cope with dynamic data in a time sequence. RNN is widely used for contextual applications such as speech because it has an internal memory to process temporal input sequences. One study [39] showed that RNN is a powerful model for sequential data and that LSTM has good performance for phoneme recognition. Sundermeyer *et al.* [40] showed that an LSTM network provided better performance than a standard RNN for an English and a large French language modeling task.

Xu et al. proposed an LSTM-CNN architecture for face antispoofing [41]. This LSTM-CNN model can learn temporal features by using a face antispoofing database with diverse attacks [42]. This architecture benefits from the combination of the CNN and the LSTM, and experimental results show that it works well for face antispoofing. However, this model focuses on face antispoofing applications. It can only be trained and used on one type of dataset.

In literature [43], a multimodal emotion recognition with evolutionary computation is proposed for human-robot interaction. This method consists of several intelligent algorithms, and the maximal recognition rate is 97%. The emotion-based communication system on the small-sized humanoid robot is practical, and this work can also be adopted in many human-robot interaction applications in the future.

In this study, the proposed method can be pretrained on a specific dataset. Previous knowledge can also be transferred and reused for the next training step by the transfer learning technique.

C. TRANSFER LEARNING

Transfer learning is a very important technology in the field of artificial intelligence (AI). With technological developments in transfer learning, the training dilemma that was originally caused by insufficient data or uneven distribution can be greatly improved. Therefore, in recent years, transfer learning has become one of the key issues in AI technology development. In this study, transfer learning technology is applied to transfer the knowledge originally learned from a large number of static images to a smaller number of dynamic images, thereby improving the AI recognition success rate. This study is practical, and experimental results demonstrate that the problem of insufficient data is solved successfully.

The transfer learning method [29] transfers knowledge of known related source data to target data. It is aimed at improving the effectiveness of the training procedure and the model performance. This learning technique also helps researchers expend lesser effort in collecting training data and takes a shorter time to train the model.

For achieving better learning performance, the basic concept of the transfer learning method is to transfer knowledge from the source domain to the target domain. However, in transfer learning, the source domain must have a strong correlation with the target domain. If two unrelated domains are transferred forcibly, the expected good results may not be obtained and the learning performance may also be reduced. With developments in AI technology, researchers hope to apply the concept of transfer learning to machine learning. Traditional machine learning, especially supervised learning, has strict requirements on the quantity of data samples, uniformity of data distribution, and complete labeling. Transfer learning solves the problem of insufficient samples and incomplete labeling in machine learning tasks by promoting efficient learning through the use of external distributed data.

Researchers have applied transfer learning to many tasks. Girshick *et al.* [44] reported that it significantly boosted the performance of object detection and segmentation. Girshick *et al.* [44] predicted poverty in the developing world by using transfer learning due to the scarce labeled data.

III. PROPOSED MODEL COMBINING CNN AND LSTM

The proposed model is aimed at learning the relation between the image sequences of human expressions and their corresponding labels. As mentioned above, facial expressions are produced by a combination of the contraction and relaxation of one or more facial muscles; therefore, they possess both appearance and temporal features.

A. CNN MODEL

First, CNN is used to capture appearance features because it provides state-of-the-art performance for several vision tasks. Fig. 1 shows the network structure. The inputs are the cropped region of interest of the image, which is also the region of the detected face. The cropped region is converted to gray scale and resized to 128 × 128 pixels. Color information is considered less relevant in facial expressions; therefore, it is not necessary to use RGB images. To reduce the memory usage, grayscale images are used in this study. The width and length of the input are 128 pixels, which is large enough for facial expressions because the face region in a 640 × 480 frame is approximately 128 × 128 pixels.

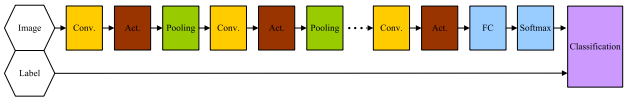


FIGURE 1. Structure of CNN.

In convolutional layer 1, the number of kernel maps is 16; size of the kernel map is 7 × 7 with the same padding method; stride length is 1; and size of feature maps is 128 × 128 × 16, that is, the width and length are the same as those of the input shape. After 2 × 2 pooling operation with 2 × 2 stride and the same padding method, the size of the feature maps becomes 64 × 64 × 16. Then, the subsequent convolutional layers contain 16, 32, 64, 64, and 128 filters with size of 7 × 7, 5 × 5, 5 × 5, 3 × 3, and 3 × 3, respectively.

After the convolutional part, feature maps with size of 4 × 4 × 128 are obtained. Finally, the feature maps are vectorized to a size of 2048 × 1 and fed to a fully connected layer. To produce features with appropriate length, we have to design the stride length of convolutional layers and pooling layers, number of kernel maps, and feasible network structure.

B. COMBINATION OF CNN AND LSTM

It is difficult for LSTMs to learn such high-dimensional data. The input image has size of 128 × 128 pixels, and it is vectorized to a size of 16384 × 1, as shown in Fig. 2. The vector is fed directly to the two-layer LSTM network with 256 cells. Then, the number of parameters in the first

LSTM layer is 17,049,600; this is too large to train the LSTM model well and efficiently.

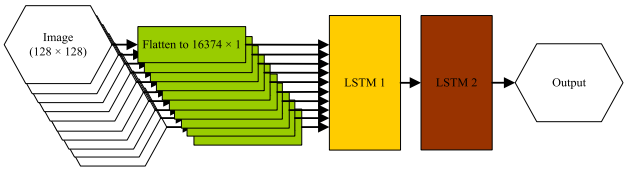


FIGURE 2. Structure of LSTMs.

Therefore, the solution is to take advantage of the CNN because it can subsample a high-dimensional image without losing important information. We use a few previous layers of the CNN as the feature extractor, as shown in Fig. 3. The extracted feature maps are flattened to create a feature vector.

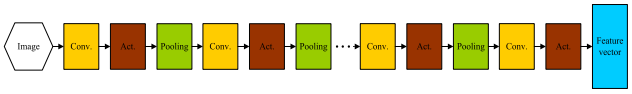


FIGURE 3. Illustration of CNN feature extractor.

The convolution layers in front of the fully connected layers of the CNN are cut and concatenated to two LSTM layers with 256 cells. Therefore, the 128 × 128 pixel input image is scaled down to a 2048 × 1 feature vector after the convolutional part. The vector is fed to the LSTM layers as the input. Then, the number of parameters in the first LSTM layer is 2,360,320; this is dramatically lower than in the case of the LSTM network without a feature extractor.

C. TRANSFERRING PARAMETERS OF CNN

To improve the model performance, we attempt to transfer the knowledge of parameters from the source domain data, as shown in Fig. 4. However, we cannot use the source data directly in the target model. First, the CNN is trained on the source task with a large number of labeled static images to enable it to extract visual information. Next, the last layer of the CNN is removed and flattened into a one-dimensional feature map. This one-dimensional feature map that is output through a model trained by source domain data is important for conducting transfer learning. This one-dimensional feature map is also imported into LSTM as time sequence information. The last layer of the CNN is removed to connect with the LSTM for target domain data training. In this way, the CNN model trained in advance with source domain data can transfer knowledge to the new model and improve the performance of the model. In this study, the source domain data is a series of static images, and the target domain data is a series of dynamic continuous images. Through transfer learning, we can transfer the knowledge of human facial emotion recognition of the static picture to the new model without having to relearn all the information. This is also a major contribution of transfer learning technology to the knowledge transfer process. Compared with the trained CNN, the untrained model with random initial weights does not have enough ability to extract useful and meaningful features.

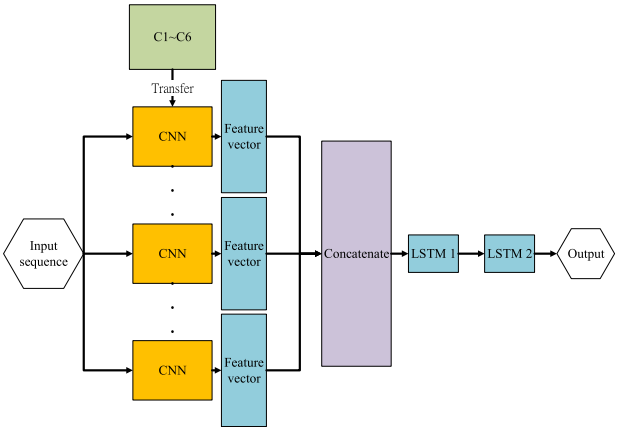


FIGURE 4. Proposed model with knowledge transfer.

With regard to facial emotion prediction, if the input uses an image from a time series, it is easy to face the problem in which the expression of the previous time point is different from that of the next time point. Because the two images are of the same subject and the time points are very close, the differences between the two images will be very small; this may cause a prediction error. In addition, if the amount of training data is seriously insufficient, it may make the model completely unable to identify images. In another situation, the expressions of the two image sequences may be significantly different; however, they are both from the same subject and the background environment is similar. In this case, when the amount of training data is insufficient, it is easy to cause misjudgment in the model. To solve these problems, the proposed model architecture adopts transfer learning technology. The proposed architecture can perform the first stage of CNN training with more source domain data. Because there is more training data, at this stage, the recognition success rate using static images as the input will be higher. Accordingly, we use the CNN with the higher recognition success rate as the feature extractor. The extracted features will be significant because the feature extractor has been trained sufficiently. In dynamic time-sensitive images, the most important image information can be inputted into the LSTM through the CNN feature extractor. This can enhance the recognition success rate and fully solve the abovementioned problems even when the target domain data is insufficient.

D. ENHANCED MODEL

After ensuring that combining the CNN and LSTM is feasible for analyzing the relations between image sequences and facial expressions, the next step is to further improve the model performance by making the neural network deeper. However, blindly increasing the number of layers is not a feasible way to enhance the performance. Thus, the residual network [46] is used to replace the convolutional part mentioned in Section III.A. A residual network is constructed using residual blocks as shown in Fig. 5(a) or using the residual

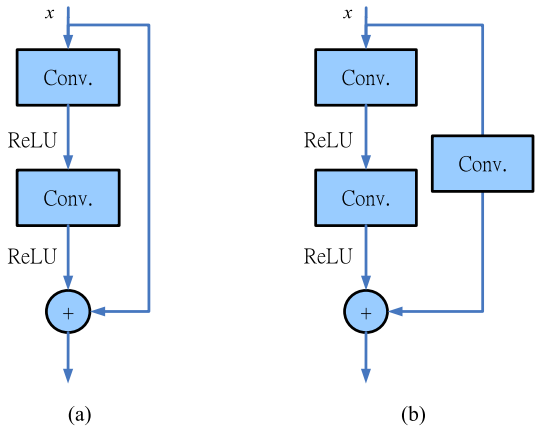


FIGURE 5. Residual block: (a) original residual block and (b) residual block with convolutional shortcut.

block with a convolutional shortcut as shown in Fig. 5(b). The convolutional shortcut makes the output shape of each path the same so that addition can be performed. The convolutional shortcut is necessary when the input and output shapes are different. Table 1 lists the model settings, and Fig. 6 shows the structure of the residual network model. In this study, the residual network was selected because the residual blocks in this architecture solve the gradient vanishing problem that often occurs during neural network training. Greater depth of the neural network increases its recognition rate, but too great a depth may cause vanishing gradient problems. Therefore, using shortcuts to connect various layers can ensure that originally distant layers have a degree of influence on each other and also maintain the gradient during the backpropagation process. Because of the special structure in the residual network, the feature extraction ability of this networ k is superior to that of the general neural network; the residual network is thus more suitable for application in this study. An LSTM unit is composed of an input gate, forget gate, cell, and output gate so that LSTM can store the important information in the cell, and this information can be written and read similarly to computer memory. The gate is designed to control the amount of information that passes through and is composed of a sigmoid layer and multiplication operation. The output of a sigmoid function is between 0 and 1 and is multiplied by the received information. The percentage of the amount of information that passes through the gate is decided. The purpose of the input gate is to decide the amount of input data to be written to the memory, which is dependent on whether the current incoming input data are important. The purpose of the forget gate is to make a decision regarding whether to keep or eliminate the previous cell state. The output gate decides whether the memory can be read. Finally, the cell updates according to the input gate and forget gate. LSTM favorably handles the time sequence problem. Therefore, the combination of favorable feature extraction by the residual network and the LSTM network is the key feature of the transfer learning method. In this study,

TABLE 1. Settings of proposed model.

Type	Settings		Parameters	
Input layer	128×128×10		N/A	
CNN Part	Conv. 32,7, 1, ReLU Max pooling		1600	
	Conv. 32,7, 1, ReLU Conv. 32,7, 1, ReLU	Conv. 32,7,1,ReLU	50208 50208	50208
	Max pooling			
	Conv. 64,3, 1, ReLU Conv. 64,3, 1, ReLU	Conv. 64,3,1,ReLU	18496 36928	18496
	Max pooling			
	Conv. 64,3, 1, ReLU Conv. 64,3, 1, ReLU	shortcut	36928 36928	N/A
	Max pooling			
	Conv. 64,3, 1, ReLU Conv. 64,3, 1, ReLU	shortcut	36928 36928	N/A
	Max pooling			
	Conv. 64,3, 1, ReLU Conv. 64,3, 1, ReLU	shortcut	36928 36928	N/A
	Conv. 128,3, 1, ReLU Conv. 128,3, 1, ReLU	Conv. 128, 3, 1, ReLU	73856 147584	73856
	Conv. 128,3, 1, ReLU Conv. 128,3, 1, ReLU	shortcut	73856 73856	N/A
	Conv. 128, 3, 1, ReLU Conv. 128, 3, 1, ReLU	shortcut	73856 73856	N/A
	Flatten			
LSTM	Output shape: 256×10 Number of cells: 256		2360320	
LSTM	Output shape: 256 Number of cells: 256		525312	
Dense	6		1542	

all the mentioned models are implemented using the Keras framework [47].

IV. EXPERIMENTAL RESULTS

The model performance is demonstrated through the results of leave-one-out cross-validation between different models. Experiments are performed using the humanoid robot Harley, and the results demonstrate the performance of the proposed real-time emotion recognition system.

A. DATABASE

1) AFFECTNET DATABASE [48]

As shown in Fig. 7, the AffectNet Database contains approximately 450,000 manually annotated and approximately 500,000 automatically annotated color images of various sizes. They are labeled as neutral, happy, sad, surprise,

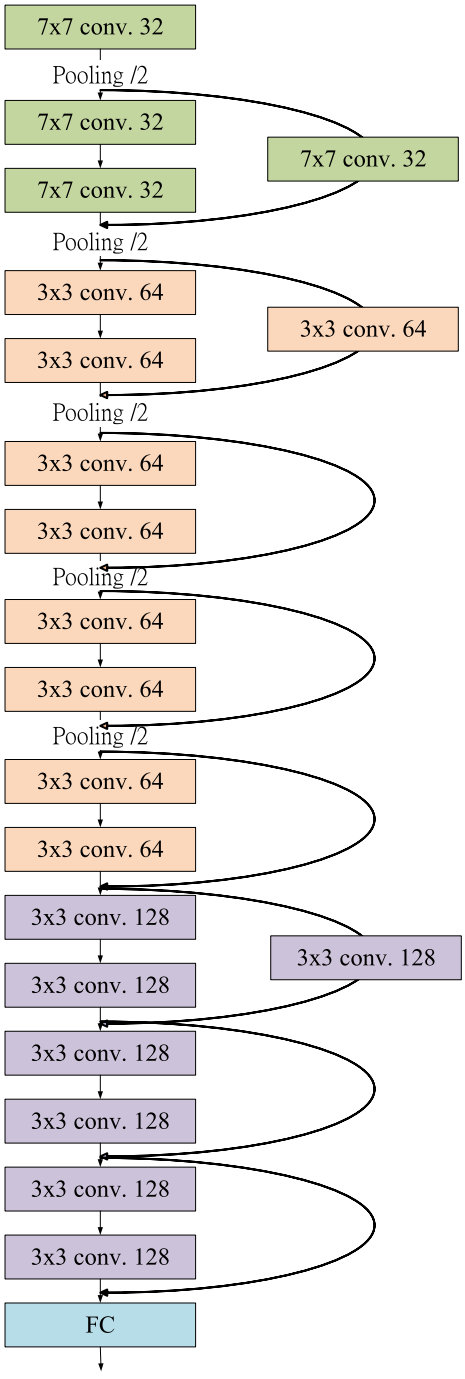


FIGURE 6. Structure of residual network.

fear, disgust, anger, and contempt. However, the dataset is highly imbalanced; there are more than 1 00,000 images with happy expressions but less than 5,000 images with disgust expressions. Therefore, we randomly selected approximately 50,000 images with six basic emotions. This includes 10,000 images each with happy, angry, surprise, and sad expressions, approximately 6000 images with disgust expression, and 3500 images with fear expression; the number of images for the last two expressions is considered insufficient.

FIGURE 7. Example of AffectNet Database [48].

2) CK+ DATABASE [49]

The extended Cohn-Kanade (CK+) dataset includes 123 subjects and 593 image sequences. Image sequences are captured using a camera located in front of the subject. All images are 640 × 480 grayscale images. The subjects are mostly Euro-American females. Among the 593 image sequences, 327 are labeled with seven expressions: happy, sad, surprise, fear, disgust, anger, and contempt. The sequences have different lengths, and most contain at least 10 images. We select 232 sequences containing more than ten images each and with six basic emotions. The last ten images of the 232 sequences are adopted in the experiments. Fig. 8 shows an example of the CK+ database. Each sequence shows the transformation in facial expressions from a neutral expression to another expression.

FIGURE 8. Example of the CK+ database. In this sequence, the subject changes expression from neutral to happy across 10 frames [49].

B. DATA PREPROCESSING

Before training the emotion recognition model, we detect the face in all static images and image sequences. The face is detected using Haar-cascade frontal face detection. Then, the detected face region is resized to 128 × 128 pixels and converted to a grayscale image. Finally, for normalization, every pixel value is divided by 255.

C. LEAVE-ONE-OUT CROSS-VALIDATION

To verify the feasibility and generalization of our proposed model, we use leave-one-out cross-validation to evaluate our proposed model and compare it with other models, such as Multilayer Perceptron (MLP), CNN, LSTM, and CNN-LSTM, that have the same structure as our proposed model but without knowledge transfer. For fairness, the settings, structures, and numbers of parameters should be as similar as possible. Tables 2–5 show the model settings.

TABLE 2. Settings of MLP.

Type	Settings	Parameters
Input layer	128×128×10	N/A
Dense 1	Batch input shape: 163840 Number of nodes: 128 Act.: sigmoid	20971648
Dense 2	Number of nodes: 128 Act.: sigmoid	16512
Dense 3	Number of nodes: 128 Act.: sigmoid	16512
Dense 4	Number of nodes: 6 Act.: Softmax	774

TABLE 3. Settings of CNN.

Type	Settings	Parameters
Conv3D	Batch input shape: 128 × 128 × 10 Number of filters:16 Filter size: 7 Stride: 1 Act.: ReLU	5504
Max pooling	Output shape: 128 × 128 × 10 × 16 Size:2 × 2 Stride:2 × 2	N/A
Conv3D	Output shape: 64 × 64 × 5 × 16 16,5,1,ReLU	32016
Max pooling	2,2,32 × 32 × 3 × 16	N/A
Conv3D	32,5,1,ReLU	64032
Max pooling	2,2,16 × 16 × 2 × 32	N/A
Conv3D	64,3,1,ReLU	55360
Max pooling	2,2,8 × 8 × 1 × 64	N/A
Conv3D	64,3,1,ReLU	110656
Max pooling	2,2,4 × 4 × 1 × 64	N/A
Conv3D	128,3,1,ReLU,4 × 4 × 1 × 128	221312
FC	Number of nodes: 100	204900
Dense	Number of nodes: 6 Act.: Softmax	606

Table 6 shows the training settings of each model. The learning rate, loss function, and optimizer are the same for all models; these three settings are the main factors influencing the learning process. CNN-LSTM uses the same structure as the proposed model but without transfer learning. The loss function and optimizer greatly influence the gradient calculation. The learning rate determines the extent to which parameters are updated with respect to the gradient.

Table 7 shows the leave-one-out cross-validation results. The second column shows the number of success in 232 validations. The model performance is clearly boosted from 58.62% (CNN-LSTM) to 90.51%, indicating the success of transfer learning. The results presented in Table 7

TABLE 4. Settings of LSTM.

Type	Settings	Parameters
Input layer	128 × 128 × 10	N/A
Flatten	Output shape: 16384 × 10	N/A
LSTM	Output shape: 256 × 10 Number of cells: 256	17040384
LSTM	Output shape: 256 Number of cells: 256	525312
Dense	6	1542

TABLE 5. Settings of CNN-LSTM.

Type		Settings	Parameters
Input layer		128 × 128 × 10	N/A
Convolution Part	Conv. Max pooling Conv. Max pooling Conv. Max pooling Conv. Max pooling Conv. Flatten	16,7, 1, ReLU 2,2 16,5, 1, ReLU 2,2 32,5, 1, ReLU 2,2 64,3, 1, ReLU 2,2 64,3, 1, ReLU 2,2 128,3,1 Output shape: 2048 × 10	149328
LSTM		Output shape: 256 × 10 Number of cells: 256	2360320
LSTM		Output shape: 256 Number of cells: 256	525312
Dense		6	1542

TABLE 6. Comparison between training settings of each model.

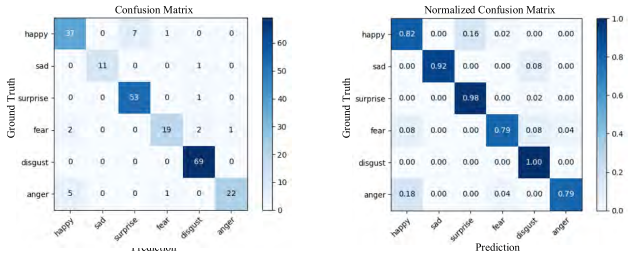
Methods	Learning rate	Epoch	Loss function	Optimizer	Parameters
MLP	0.0001	100	Cross entropy	Adam [50]	21M
CNN	0.0001	100	Cross entropy	Adam	690K
LSTM	0.0001	100	Cross entropy	Adam	17M
CNN-LSTM	0.0001	100	Cross entropy	Adam	3M
Proposed model	0.0001	100	Cross entropy	Adam	4M

TABLE 7. Comparisons between different models on CK+ database with leave-one-out cross-validation.

Methods	Number of successes	Accuracy (%)
MLP	94	40.5
CNN	158	68.1
LSTM	148	63.79
CNN-LSTM	136	58.62
Proposed model	210	90.51

also indicate that the traditional MLP architecture has poor recognition ability (only 40.5% accuracy) with time sequence information, whereas the general CNN (68.1% accuracy) and LSTM (63.79% accuracy) have slightly superior ability. However, if transfer learning and pretraining are not implemented in the first half of the same architecture, the accuracy is only 58.62%. After transfer learning and CNN architecture pretraining, the recognition ability improved to 90.51% accuracy. Although a deep neural network architecture is critical, recognition performance improvement through transfer learning is even more vital.

The confusion matrix of the leave-one-out cross-validation experiment is displayed in Fig. 9. The numbers in Fig. 9(a) are the actual tested data of the experiment. However, because the data distribution is not uniform for various emotions, the normalized confusion matrix is presented in Fig. 9(b). The diagonal line of the confusion matrix is the ratio of the predicted result to the actual result. Fig. 9(b) reveals that the darker color is concentrated along the diagonal, which confirms that the accuracy of the model proposed in this paper is high. The recognition rates for happiness, fear, and anger were 0.82, 0.79, and 0.79, respectively; in addition, the scores for sadness and surprise were relatively high (0.92 and 0.98). A 100% accuracy rate was achieved for surprise. In the actual data for happiness, seven instances were misjudged as surprise and one was misjudged as fear. In the actual data for fear, two instances are misjudged as happiness and two are misjudged as disgust. In the actual data for anger, five instances were misjudged as happiness and one was misjudged as fear. When these numbers are examined in detail, part of the reason for the generally low accuracy rate is clearly that the number of samples was low for the emotions with lower accuracy rate, resulting in insufficient training and testing data. Nonetheless, the numbers of misjudgments were within an acceptable range. The performance was relatively high for the other emotion categories in which more samples were used.



(a) (b)

FIGURE 9. Confusion matrix in terms of (a) the test number and (b) the normalized value.

Furthermore, the contingency table of six emotions are illustrated in Table 8-13. All occurred numbers in true positive, false negative, false positive, and true negative are completely addressed in these tables. For all emotions, the true positive and true negative values denote the numbers of the correct prediction results. Obviously, the obtained values of true positive and true negative in each table are much higher than the numbers of false positive and false negative. It also demonstrates the performance and the stability of the proposed approach. On the other hand, the false positive and false negative values represent the incorrect prediction results. As shown in Table 8-13, all the false positive and false negative values are extremely small in all the experiments.

TABLE 8. Contingency table of happy.

		Prediction
Ground Truth	37 (True Positive)	8 (False Negative)
	7 (False Positive)	180 (True Negative)

TABLE 9. Contingency table of sad.

		Prediction
Ground Truth	11 (True Positive)	1 (False Negative)
	0 (False Positive)	220 (True Negative)

TABLE 10. Contingency table of surprise.

		Prediction
Ground Truth	53 (True Positive)	1 (False Negative)
	7 (False Positive)	168 (True Negative)

TABLE 11. Contingency table of fear.

		Prediction
Ground Truth	19 (True Positive)	5 (False Negative)
	2 (False Positive)	206 (True Negative)

TABLE 12. Contingency table of disgust.

		Prediction
Ground Truth	69 (True Positive)	0 (False Negative)
	4 (False Positive)	159 (True Negative)

TABLE 13. Contingency table of anger.

		Prediction
Ground Truth	22 (True Positive)	6 (False Negative)
	1 (False Positive)	203 (True Negative)

In spite of the fact that the obtained false positive and false negative values are slightly higher in Table 8, the true positive and the true negative values are still much higher than the false positive and false negative values. It also shows the prediction errors are acceptable for recognizing the facial emotions.

After transferring the layer and parameters, the different training conditions are compared to determine how to train the model well. Fig. 10 shows the different training conditions. Square and circle symbols represent the convolutional layer and the residual block, respectively. Green squares and circles are trainable and blue ones are fixed, nontrainable, or frozen, indicating that the parameters in the residual block or convolutional layer will not be updated. The first one on the left in Fig. 10 is to fine-tune the whole model after transferring. The last one is to freeze the whole transferred part so that it can be treated as a fixed feature extractor. The number of parameters decreases from left to right.

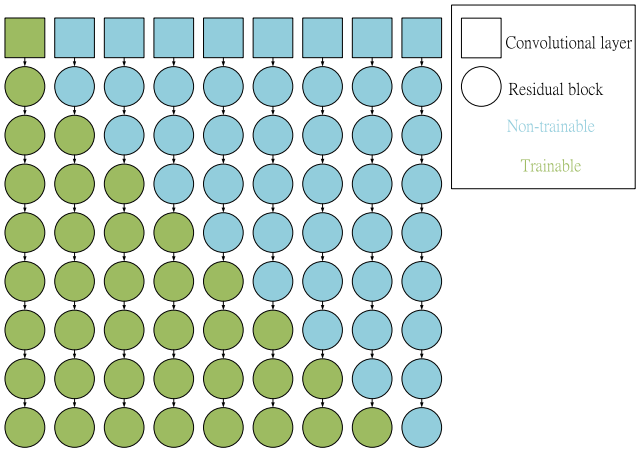


FIGURE 10. Illustration of training condition.

The comparison results in Fig. 11. show the training loss, training accuracy, validation loss, and validation accuracy of the nine conditions including fine-tune training (FTT, first one on left in Fig. 10) and F1T-F8T (FnT means that the first n residual block are frozen and the rest of the network is trained). In this simulation, the CK+ database is used as training and validation data, and the ratio of validation data to training data is set to 0.1. In other words, 10% of the 232 sequences are randomly selected as validation data. Furthermore, 50 training epochs are used and the training process is repeated 10 times.

In Fig. 11(a) and 11(b), the training process of FTT is significantly better than that of freeze training but is similar to that of the model with only a few frozen residual blocks. However, whether th e training process of the model is good is not the only evaluation indicator. Therefore, the validation is a helpful way to verify the performance and generalization ability of the model. In Fig. 11(c) and 11(d), FTT clearly has the lowest validation loss and the highest validation accuracy compared with other approaches.

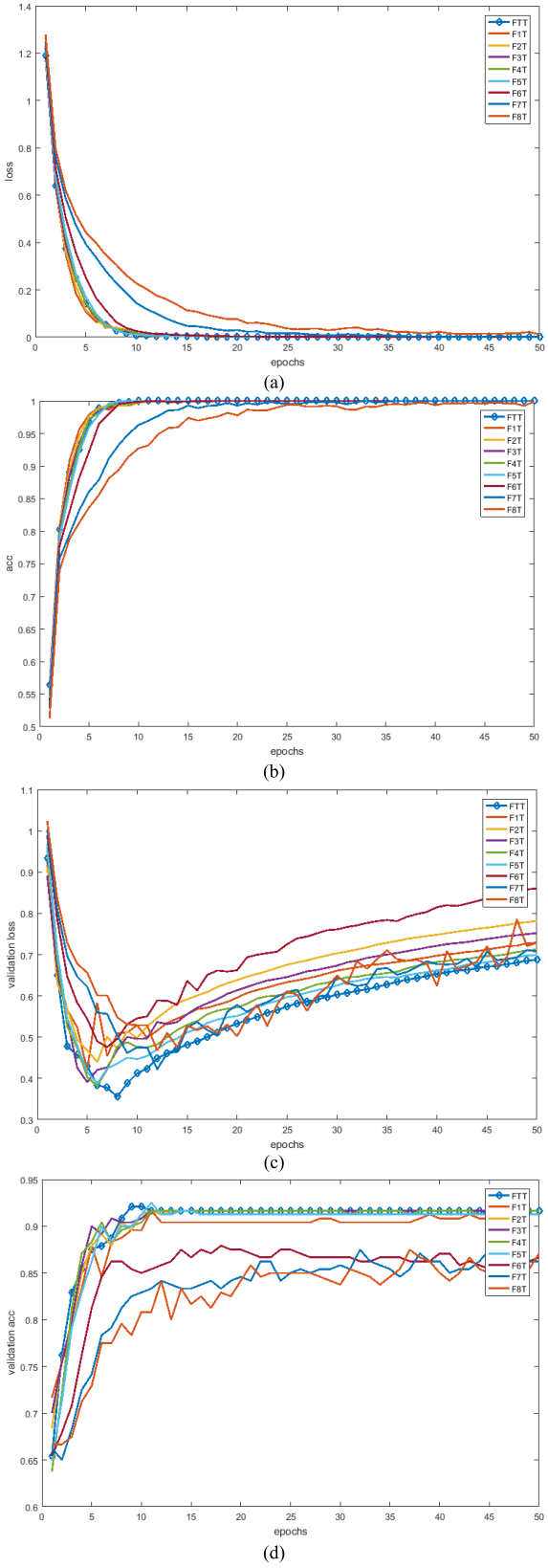
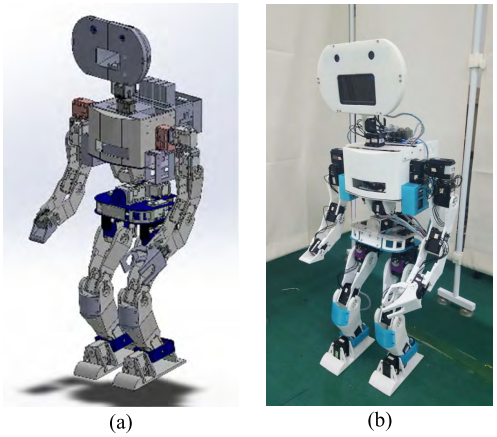


FIGURE 11. Comparison between freeze and fine-tune: (a) training loss, (b) training accuracy, (c) validation loss, and (d) validation accuracy.

In the transfer learning process, knowledge of the source domain data is stored in the CNN architecture, and the one-dimensional feature map generated by the CNN model serves as the input of the LSTM. Therefore, at this moment, the internal parameters of the CNN have already been trained but the LSTM is still in the initial state that has not been trained. For this reason, when the CNN and the LSTM are linked together for training, whether or not the CNN parameters need to be retrained or be kept unchanged will be a very important issue. This experiment mainly explores the impact of whether CNN undergoes partial training on the transfer learning results. It is also a type of neural network hyperparameter that determines whether it is necessary to retrain all or only a few layers of the CNN. This hyperparameter has an extremely significant impact on transfer learning. However, the parameter adjustments in the CNN and LSTM have less impact. The experiment results suggest that applying transfer learning in facial expression recognition, performing feature extraction with a pretrained CNN, inputting the extracted feature to the LSTM, and finally optimizing the CNN's internal parameters is a better approach.

D. EXPERIMENTAL SETUP

An emotion recognition experiment is conducted using the humanoid robot Harley (Fig. 12), which is inspired by childhood developmental milestones. We apply the emotion recognition system to Harley to enable it to perform basic emotion re cognition and interact with humans.



(a) (b)

FIGURE 12. Harley humanoid robot: (a) mechanism design and (b) photograph.

E. EXPERIMENT I

In this experiment, the emotion recognition model is trained on the CK+ database. First, the camera is used to capture images of users. Then, data preprocessing is performed on these images. The resized cropped grayscale image is stored in a buffer until the buffer is filled with 10 frames. Then, image sequences stored in the buffer are predicted by the trained emotion recognition model. Finally, the results of the predicted emotion and the corresponding probability are shown. Fig. 13 shows the experimental results for six subjects.

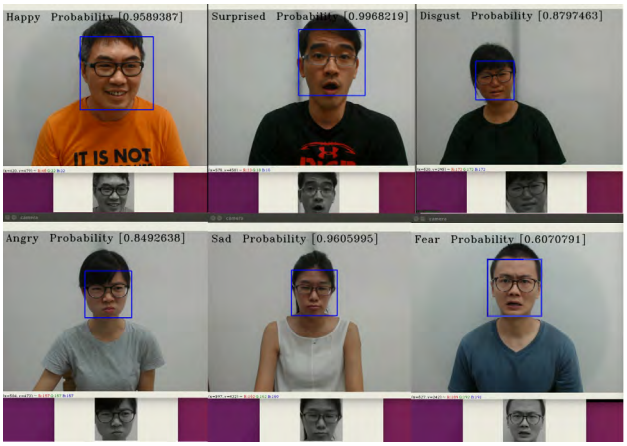


FIGURE 13. Experimental result of emotion recognition.

F. EXPERIMENT II

The proposed emotion recognition system is aimed at improving the HRI. In this experiment, the Harley robot is used to evaluate subjects' emotion states from their facial expressions. Fig. 14 shows the environment setup, and Fig. 15 depicts the interaction between Harley and user.

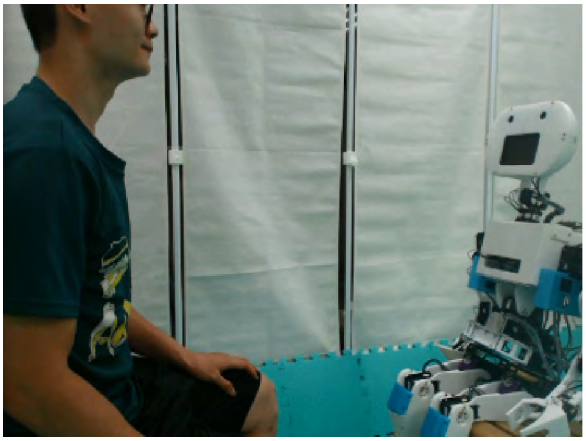


FIGURE 14. Experimental environment.

First, the user and Harley sit face-to-face, and Harley greets the user upon detecting their face. The user can articulate, express feelings, or say something. Harley evaluates the user's emotions by collecting their facial images and predict the emotion from these images using the proposed model. Harley then provides an appropriate response to the user. It should be noted that Harley does not possess a dialogue system and that the proposed emotion recognition system is based on facial expression analysis rather than semantic analysis. To clearly demonstrate the feasibility and practicality of the emotion recognition application, the whole experimental video is uploaded; it can be accessed in [51].

V. DISCUSSIONS

In recent years, many related papers on facial emotion recognition have been published. One study [52] used an extreme

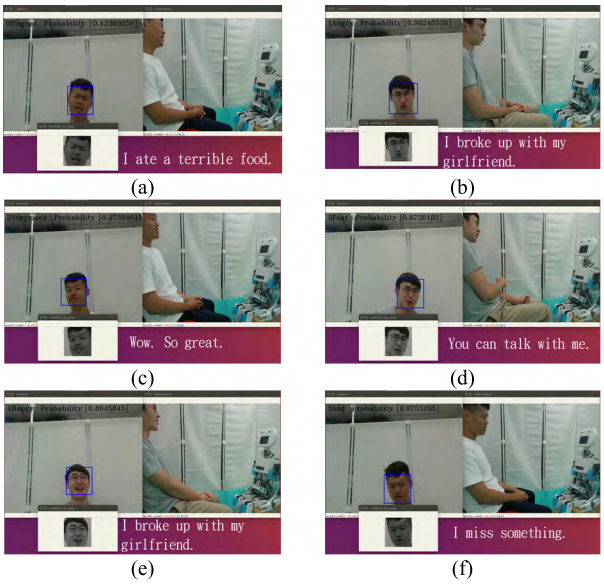


FIGURE 15. Interaction between Harley and user: (a) disgust, (b) angry, (c) surprised, (d) fear, (e) happy, and (f) sad.

learning machine (ELM) with optical flow to achieve this goal and reported good performance in experiments. However, because of the different characteristics of the various data sets, the recognition rate was also widely distributed (from 29.81% to 61.8% to 92.74%). Therefore, the characteristic of the data set has a very large impact on the model performance. Another study [53] used the constrained local model (CLM) and key emotion (KE) points for facial emotion prediction. The experiment results suggest that this method provides good performance for different datasets as indicated by the true positive rate (TPR) parameter used in this study. However, this method must successfully capture two neural frames to build the model. If these two neural frames are captured improperly, it will affect the method's subsequent performance. A study [54] applied hybrid deep neural networks for continuous facial recognition. This method combined a CNN comprising six simple convolutional layers and a general-type RNN. The experiment results indicate that under the Japanese Female Facial Expression (JAFFE) database [55] test, the highest accuracy of 94.91% can be achieved. The accuracy of this method is good but is still limited by the dataset. The highest recognition of facial emotions can only be so effective in the expression of Japanese women. Another study [56] applied the optical flow method to acquire the feature vector and input the feature vectors into a neural network for facial emotion recognition. This method combined traditional image processing methods with machine learning techniques and achieved good validation in experiments. Two different datasets were used in the experiments for performance verification, and accuracy rates of 75.25% and 70.93% were achieved, respectively.

The abovementioned technologies are the research results of facial emotion recognition in recent years. The datasets used by these methods are not the same, and each of these models have their own advantages and disadvantages. This study focuses on how to use transfer learning to further improve the recognition success rate from 58.62% to 90.51%, and it provides a detailed discussion of the training process for knowledge transfer. In the future, this technology can be extended to various fields to enhance the recognition rate of datasets with a limited amount of data. In addition, in the analysis of computing time, the above methods all have the ability to detect emotions in real-time. This study also fully demonstrates the performance of real-time computing, as shown in the video of Experiment II [51]. Therefore, in terms of computing time, the proposed method has superior performance.

With regard to human facial emotion recognition, many related studies have integrated audio and image data for analysis. One study [57] applied the hybrid deep model that first inputs image and audio signals into the audio network and visual network, both of which use the CNN architecture. Finally, the emotion predictions are output through Restricted Boltzmann machines (RBMs) and support vector machines (SVM). Among the results of the various dataset tests, the highest accuracy rate achieved was 85.97%. The approach used in [58] was to preprocess images and sounds. The images first underwent grayscale conversion, and the audio underwent framing windowing and Fourier transform. Emotion recognition is performed through ELM and SVM. Accuracy rate of 86. 4% was achieved for the eNTERFACE database [59]. One study [60] considered both images and audio for facial emotion recognition. This method combines principal component analysis (PCA) with other commonly used machine learning algorithms for emotion prediction. In addition, the article also lists the effects of various different algorithms on experimental results. The experimental results show that the method combining random forest (RF) with PCA achieves the best performance for the eNTERFACE database. Other studies, [57] and [61] also used the CNN architecture to separately extract features for audios and images and finally output the emotion prediction results through RBMs and SVM. Both methods provided good verification results in the experiments. However, because the dimension of the image data is very large, dimensionality reduction must be performed. Therefore, one study [62] adopted bi-directional principal component analysis (BDPCA) and least-squares linear discriminant analysis (LSLDA) for solving this issue. Further, the extracted features were inputted into the optimized kernel-Laplacian radial basis function (OKL-RBF) neural classifier. Audio data is analyzed with prosodic features and Mel-scale frequency cepstral coefficients. One study [62] combined the abovementioned processes to achieve average recognition rate of 86.67%. The above research results for audio-visual emotion recognition were obtained in recent years. According to one of these studies [58], the size of the eNTERFACE database is still insufficient. Therefore, this database can achieve limited improvement in recognition accuracy. This further confirms the importance of transfer learning as described in this paper.

In the future, this concept can be extended to various machine learning applications, and the performance of AI can be improved further.

In addition, facial expression recognition methods that involve machine learning include SVM, as described in the literature [63]. This method combines geometric deformation features with SVM and applies it to recognize expressions in image sequences. Its performance has been well validated in experiments. Regarding speech emotion recognition, one study [64] proposed an anchor model to solve this problem, and the Mel frequency cepstral coefficient (MFCC) and Gaussian mixture model (GMM) techniques have also been employed. One study [65] applied the MFCC technique and added the Fourier parameters as a consideration factor for speech emotion recognition. Another study [66] employed SVM for speech emotion recognition, and the performance of the method was verified using the Geneva Whispered Emotion Corpus database. The aforementioned speech emotion recognition methods have not been tested using the same database, but the experimental results confirm that these methods achieve favorable performance.

Facial AUs, geometric features, and graph-based modelling techniques are generally applied separately in facial expression analysis. However, Ghayoumi and Bansal [67] united geometric features with facial AUs by employing PCA and SVM. The AUs are mapped to the geometric features in the procedure. In experiments, the six basic emotions were used in the approach. The performance improved by 70%.

Because vision-based facial expression analysis has limited accuracy, the difficulty of its recognition process is high. However, in the real environment, all the required information in any situation cannot be obtained. Furthermore, if only a small amount of the vision information can be acquired, the recognition difficulty is considerably higher. Attempting to collect all the useful information to use in the multimodal approach is one favorable solution to this problem. However, this article aimed to solve the problem of having limited training data, and the transfer learning concept was employed in the training process of the proposed model. The experiments detailed herein reveal that the transfer learning method is feasible and practical to use for solving this problem.

This study proposed a transfer learning approach for solving the lack of large amount data problem. The experimental results show that the transfer learning concept can help the learning system to obtain higher performance. The solution is expectedly sub-optimal, and it could be also possibly outperformed by other approaches in a data abundance scenario.

VI. CONCLUSION

This study proposes an emotion recognition system based on a deep neural network for improving the HRI. This system is applied to a humanoid robot. Users' facial images are captured using a camera mounted on the head of a humanoid robot. The robot then provides appropriate responses to the user according to their emotions as recognized using the proposed model. The proposed model combines CNN and LSTM and exploits the advantages of both CNN and RNN. Leave-one-out cross-validation indicates that the model performance is improved significantly. The feasibility and practicability of this model are validated.

APPENDIX

See Table 14.

TABLE 14. Table of acronyms.

Acronym	Explanation
AU	Action Unit
AI	Artificial Intelligence
BDPCA	Bi-Directional Principal Component Analysis
CLM	Constrained Local Model
CNN	Convolutional Neural Network
FACS	Facial Action Coding System
GMM	Gaussian Mixture Model
HRI	Human–Robot Interaction
JAFFE	Japanese Female Facial Expression
KE	Key Emotion
LSLDA	Least-Squares Linear Discriminant Analysis
LSTM	Long Short-Term Memory
MFCC	Mel Frequency Cepstral Coefficient
MLP	Multilayer Perceptron
OKL-RBF	Optimized Kernel-Laplacian Radial Basis Function
PCA	Principal Component Analysis
RBMs	Restricted Boltzmann Machines
RF	Random Forest
RNN	Recurrent Neural Network
SVM	Support Vector Machines
TPR	True Positive Rate

REFERENCES

[1] K. Dautenhahn, "Methodology & themes of human-robot interaction: A growing research field," *Int. J. Adv. Robotic Syst.*, vol. 4, no. 1, p. 15, Mar. 2007.

[2] L. E. Parker, F. E. Schneider, and A. C. Schultz, *Multi-Robot Systems: From Swarms to Intelligent Automata*. Dordrecht, The Netherlands: Springer, 2005.

[3] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Sci. Inf.*, vol. 44, no. 4, pp. 695–729, 2005.

[4] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, vol. 3. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[5] M. S. Bartlett, G. Donato, J. C. Hager, P. Ekman, J. R. Movellan, and T. J. Sejnowski, "Face image analysis for expression measurement and detection of deceit," in *Proc. Annu. Joint Symp. Neural Comput.*, 1999, pp. 8–15.

[6] M. Ghayoumi, M. Thafar, and A. K. Bansal, "Towards formal multimodal analysis of emotions for affective computing," in *Proc. Int. Conf. Distrib. Multimedia Syst.*, 2016, pp. 48–54.

[7] R. Plutchik and H. Kellerman. *Theories of Emotion*. New York, NY, USA: Academic, 2013.

[8] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. K. Heider, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, K. Scherer, M. Tomita, and A. Tzavaras, "Universals and cultural differences in the judgments of facial expressions of emotion," *J. Personality Social Psychol.*, vol. 53, no. 4, pp. 712–717, 1987.

[9] F. T. Giuntini, L. P. Ruiz, L. De F. Kirchner, D. A. Passarelli, M. De J. D. Dos Reis, A. T. Campbell, and J. Ueyama, "How do I feel? Identifying emotional expressions on facebook reactions using clustering mechanism," *IEEE Access*, vol. 7, pp. 53909–53921, 2019.

[10] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affective Comput.*, vol. 2, no. 2, pp. 92–105, Apr. 2011.

[11] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis," *IEEE Trans. Affective Comput.*, vol. 1, no. 2, pp. 81–97, Jul. 2010.

[12] B. Yang, J. Cao, R. Ni, and Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*, vol. 6, pp. 4630–4640, 2017.
[13] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Trans. Affective Comput.*, vol. 8, no. 3, pp. 300–313, Jul./Sep. 2017.
[14] V. Mayya, R. M. Pai, and M. M. M. Pai, "Automatic facial expression recognition using DCNN," *Procedia Comput. Sci.*, vol. 93, pp. 453–461, Jan. 2016.
[15] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.
[16] Y.-H. Byeon and K.-C. Kwak, "Facial expression recognition using 3D convolutional neural network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 5, no. 12, pp. 1–6, 2014.
[17] W. Zhang, Y. Zhang, L. Ma, J. Guan, and S. Gong, "Multimodal learning for facial expression recognition," *Pattern Recognit.*, vol. 48, no. 10, pp. 3191–3202, 2015.
[18] X. Fan and T. Tjahjadi, "A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences," *Pattern Recognit.*, vol. 48, no. 11, pp. 3407–3416, Nov. 2015.
[19] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.
[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[21] L. Zou, J. Zheng, C. Miao, M. J. Mckeown, and Z. J. Wang, "3D CNN based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural MRI," *IEEE Access*, vol. 5, pp. 23626–23636, 2017.
[22] Y. Yang, J. Yue, J. Li, and Z. Yang, "Mine water inrush sources online discrimination model using fluorescence spectrum and CNN," *IEEE Access*, vol. 6, pp. 47828–47835, 2018.
[23] Q. Zheng, M. Yang, J. Yang, Q. Zhang, and X. Zhang, "Improvement of generalization ability of deep CNN via implicit regularization in two-stage training process," *IEEE Access*, vol. 6, pp. 15844–15869, 2018.
[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[25] S. Chen, B. Song, and J. Guo, "Attention alignment multimodal LSTM for fine-gained common space learning," *IEEE Access*, vol. 6, pp. 20195–20208, 2018.
[26] F. Zhang, C. Hu, Q. Yin, W. Li, H.-C. Li, and W. Hong, "Multi-aspect-aware bidirectional LSTM networks for synthetic aperture radar target recognition," *IEEE Access*, vol. 5, pp. 26880–26891, 2017.
[27] R. Zazo, P. S. Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak, "Age estimation in short speech utterances based on LSTM recurrent neural networks," *IEEE Access*, vol. 6, pp. 22524–22530, 2018.
[28] M. Sundermeyer, H. Ney, and R. Schlüter, "From feedforward to recurrent LSTM neural networks for language modeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 3, pp. 517–529, Mar. 2015.
[29] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[30] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proc. Nat. Acad. Sci. USA*, vol. 79, no. 8, pp. 2554–2558, 1982.
[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167
[33] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, Jun. 2011, pp. 315–323.
[34] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[35] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.
[36] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 435–442.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, Sep. 2014, pp. 1–4.
[38] M. Ghayoumi and A. K. Bansal, "Multimodal architecture for emotion in robots using deep learning," in *Proc. Future Technol. Conf. (FTC)*, Dec. 2016, pp. 901–907.
[39] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 6645–6649.
[40] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling," in *Proc. Interspeech*, Sep. 2012, pp. 194–197.
[41] Z. Xu, S. Li, and W. Deng, "Learning temporal features using LSTM-CNN architecture for face anti-spoofing," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit.*, Nov. 2015, pp. 141–145.
[42] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, Apr. 2012, pp. 26–31.
[43] L.-A. Perez-Gaspar, S.-O. Caballero-Morales, and F. Trujillo-Romero, "Multimodal emotion recognition with evolutionary computation for human-robot interaction," *Expert Syst. Appl.*, vol. 66, pp. 42–61, Dec. 2016.
[44] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.
[45] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science*, vol. 353, no. 6301, pp. 790–794, 2016.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[47] *Keras: The Python Deep Learning Library*. Accessed: Mar. 25, 2019. [Online]. Available: https://keras.io/
[48] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affective Comput.*, vol. 10, no. 1, pp. 18–31, Jan./Mar. 2019.
[49] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2010, pp. 94–101.
[50] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, Dec. 2014, pp. 1–15.
[51] *Experimental Video*. Accessed: Jan. 28, 2019. [Online]. Available: https://www.youtube.com/watch?v=3cQC6mC0W_k
[52] S. Shojaeilangari, W.-Y. Yau, K. Nandakumar, J. Li, and E. K. Teoh, "Robust representation and recognition of facial emotions using extreme sparse learning," *IEEE Trans. Image Process.*, vol. 24, no. 7, pp. 2140–2152, Jul. 2015.
[53] P. Chiranjeevi, V. Gopalakrishnan, and P. Moogi, "Neutral face classification using personalized appearance models for fast and robust emotion detection," *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2701–2711, Sep. 2015.
[54] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognit. Lett.*, vol. 115, pp. 101–106, Nov. 2018.
[55] M. J. Lyons, J. Budynek, and S. Akamatsu, "Automatic classification of single facial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 12, pp. 1357–1362, Dec. 1999.
[56] G. Patil and P. Suja, "Emotion recognition from 3D videos using optical flow method," in *Proc. Int. Conf. Smart Technol. Smart Nation*, Aug. 2017, pp. 825–829.
[57] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, Oct. 2018.

[58] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio-visual emotional big data," *Inf. Fusion*, vol. 49, pp. 69–78, Sep. 2019.
[59] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE' 05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2006, p. 8.
[60] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Trans. Affective Comput.*, vol. 10, no. 1, pp. 60–75, Jan./Mar. 2019.
[61] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (AVEF): A deep efficient weighted approach," *Inf. Fusion*, vol. 46, pp. 184–192, Mar. 2019.
[62] K. P. Seng, L.-M. Ang, and C. S. Ooi, "A combined rule-based & machine learning audio-visual emotion recognition approach," *IEEE Trans. Affective Comput.*, vol. 9, no. 1, pp. 3–13, Jan./Mar. 2018.
[63] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 172–187, Jan. 2007.
[64] Y. Attabi and P. Dumouchel, "Anchor models for emotion recognition from speech," *IEEE Trans. Affective Comput.*, vol. 4, no. 3, pp. 280–290, Jul. 2013.
[65] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using Fourier parameters," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 69–75, Jan. 2015.
[66] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Exploitation of phase-based features for whispered speech emotion recognition," *IEEE Access*, vol. 4, pp. 4299–4309, 2016.
[67] M. Ghayoumi and A. K. Bansal, "Unifying geometric features and facial action units for improved performance of facial expression analysis," in *Proc. Int. Conf. Circuits, Syst., Signal Process., Commun. Comput. (CSSCC)*, 2015, pp. 259–266.

TZUU-HSENG S. LI (S'85–M'90) received the B.S. degree from the Tatung Institute of Technology, Taipei, Taiwan, in 1981, and the M.S. and Ph.D. degrees from National Cheng Kung University (NCKU), Tainan, Taiwan, in 1985 and 1989, respectively, all in electrical engineering, where he has been with the Department of Electrical Engineering, since 1985 and he is currently a Distinguished Professor. From 1996 to 2009, he was also a Researcher with the Engineering and Technology Promotion Center, National Science Council, Tainan. From 1999 to 2002, he was the Director of the Electrical Laboratories, NCKU. From 2009 to 2012, he was the Dean of the College of Electrical Engineering and Computer Science, National United University, Miaoli, Taiwan. From 2009 to 2016, he was the Vice President of the Federation of International Robot-Soccer Association. Since 2014, he has been the Director of the Center for Intelligent Robotics and Automation, NCKU. His current research interests include artificial and biological intelligence and applications, fuzzy systems and control, home service robots, humanoid robots, mechatronics, 4WIS4WID vehicles, and singular perturbation methodology. He was elevated to CACS Fellow and RST Fellow, in 2008 and 2018, respectively. He was a recipient of the Outstanding Automatic Control Award from the Chinese Automatic Control Society (CACS), Taiwan, in 2006, and the Outstanding Research Award from the Ministry of Science and Technology, Taiwan, in 2017. He was a Technical Editor of the IEEE/ASME TRANSACTIONS ON MECHATRONICS and an Associate Editor of the *Asia Journal of Control*. He is currently an Editor-in-Chief of *iRobotics*, and an Associate Editors of the *International Journal of Electrical Engineering*, the *International Journal of Fuzzy Systems*, and the IEEE TRANSACTIONS ON CYBERNETICS. He was elected as the President of the CACS, from 2008 to 2011 and the Robotics Society of Taiwan, from 2012 to 2015.

PING-HUAN KUO received the B.S., M.S., and Ph.D. degrees from the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, in 2008, 2010, and 2015, respectively. Since 2017, he has been with the Computer and Intelligent Robot Program for Bachelor Degree, National Pingtung University (NPTU), where he is currently an Assistant Professor. His current research interests include fuzzy control, intelligent algorithms, humanoid robot, image processing, robotic application, big data analysis, machine learning, and deep learning applications.

TING-NAN TSAI received the B.S. and the M.S. degrees from the Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, in 2016 and 2018, respectively. His current research interests include fuzzy control, intelligent systems, humanoid robot, image processing, robotic application, and FIRA/RoboCup game.

PO-CHIEN LUAN received the B.S. degree from the Department of Electrical Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan, in 2018, where he is currently pursuing the M.S. degree. His current research interests include fuzzy control, intelligent systems, humanoid robot, image processing, robotic application, and FIRA game.