Facial expression recognition using bidirectional LSTM - CNN

Article *in* Procedia Computer Science · January 2023

DOI: 10.1016/j.procs.2022.12.109

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 216 (2023) 39–47

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Computer Science and Computational Intelligence 2022

Facial expression recognition using bidirectional LSTM - CNN

Rio Febrian[a]*, Benedic Matthew Halim[a], Maria Christina[a], Dimas Ramdhan[a], Andry Chowanda[a]

[a]Computer Science Department, School of Computer Science, Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia

Abstract

Nowadays, there has been much attention on computer vision regarding human-computer interaction, especially facial expression recognition (FER). Many researchers have explored and suggested systems for this field. In this paper, we propose the Deep Learning architecture to improve the performance of models from the previous work. Additionally, we propose the BiLSTM-CNN model, which combines our proposed CNN and BiLSTM model. Besides that, we also compare the model to our CNN and LSTM-CNN models. We conduct the experiments on the CK+ dataset and evaluate the accuracy rate of the built models. Data augmentation is used in the dataset to improve the model's performance and prevent overfitting. The results demonstrate that the BiLSTM-CNN method achieves a state-of-the-art accuracy rate compared to other methods from previous work. The highest accuracy of 99.43% is reached by the BiLSTM-CNN model with data augmentation.

Keywords: Deep Learning; CNN; LSTM; Bidirectional LSTM

1. Introduction

Humans have been using emotions to show others what they are feeling. These emotions are also used as a means of communication to relay information. Over the last few decades, there have been issues regarding mutual sympathy in human-computer interactions. But there has been much research regarding emotional communication between humans and machines. This matter has received much attention as part of human-computer interaction.

For the past two decades, the Artificial Intelligence community has explored and suggested systems for automatic emotion recognition [1]. Facial expressions, body posture, movements, linguistic structure, and environment can all

* Corresponding author. Tel.: +62-896-872-437-56.
E-mail address: rio.febrian@binus.ac.id

10.1016/j.procs.2022.12.109

indicate emotional states. In one of the most extensive investigations on human emotions, Ekman and Friesen identified six universal emotions based on facial expressions: disgust, fear, happiness, surprise, sadness, and rage.

The usage of a collection of distinct features, which constrains the solution to specified rules, is one of the most important properties of the presented systems. When employed for facial expression detection, methods based on hard-coded characteristics produce excellent results. Still, they cannot be implemented in a real-world scenario due to numerous constraints such as lighting, subject posture, and skin colour. In recent years, the use of implicit features for picture classification has successfully overcome this difficulty in the computer vision field. Deep neural networks were offered as an implicit feature model because they use the data to learn the most important parts of the image. One of the benefits of deep neural networks is their generalisation ability, as well as the low computational cost of extracting and classifying data once they've been trained. Convolutional Neural Networks (CNN) are the deep neural architectures that show the most promise for face-processing applications.

As has been stated before, facial expressions are composed of numerous constraints that are based on hard-coded characteristics; a spatial relationship is needed between different facial regions. However, CNN filters seem to be failed to capture such a relationship because they are only performed locally on image regions. Bidirectional LSTM is needed for our network model to learn from complete time series at each time step. Therefore, we need to improve our face expression recognition performance by exploring long-term dependencies, especially for the spatial dependencies within facial expression images.

2. Related Works

Facial expressions are vital in the process of exchanging emotional information between humans compared to direct speech, and these expressions can be used as evidence to see whether an individual is telling the truth or not [2,3]. The use of facial recognition technology is growing [4], for example, to detect fake faces in photos and real faces through the eyes [5]. With the development of technology, various methods have emerged to detect human facial expressions.

Initially, the facial expression recognition method has been implemented traditionally, but the result shows that it takes a very long time to extract features from images [6,7]. This is because FER has a factor in image lighting, erratic angles, low resolution, or incomplete image data. Along with the development of research, an algorithm has emerged that can extract image features optimally called Convolutional Neural Network (CNN).

Convolutional Neural Networks (CNN), which have high accuracy for face recognition, have been proven by comparison with other methods. This makes many people interested in the CNN method; moreover, apart from its high accuracy for recognising faces, CNN can also be used in real time. The accuracy of face recognition through images and in real-time is more than 90%, which means this method is optimal. However, this CNN method must be trained using multiple images to achieve optimal results [8,9]. In addition, we can also simplify the CNN model by layering the convolutional and sampling layers together. Fig. 1 shows the general architecture of CNN:

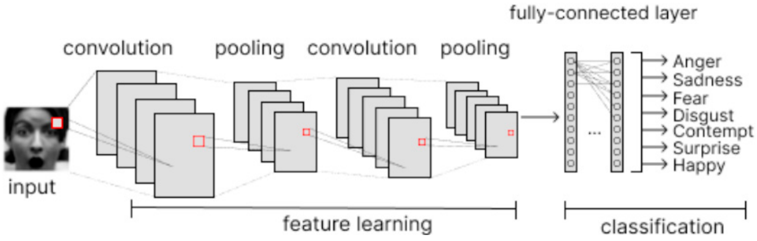


Fig. 1. General Architecture of CNN

This algorithm is also often combined with other algorithms or methods, such as those carried out in research, which combines Bidirectional LSTM [10], Residual Masking Network [11], Multi-branch Cross Connection [12], and Adaboost respectively, with CNN. This aims to improve the accuracy and performance of the CNN algorithm in detecting someone's facial expressions.

In addition to merging with CNN, several other methods can be used to detect a person's facial expression. Researchers in [13] used the concept of a tree data structure named the Minimum Spanning Tree. The Weakness in this study is the speed of classification, which is less but produces a relatively good and optimal level of detection accuracy. In addition, research conducted by [14] uses Visual Transformers Feature Fusion to classify image information well. Finally, researchers in [15] used another deep neural network, namely the Artificial Neural Network, to perform accurate filters on facial image features. Thus, the level of accuracy of facial expression detection is increasing.

The methods and algorithms that were found will certainly be implemented in the real world. For example, research in [16] used a humanoid robot as an experiment to see how the robot responds to recognizing a person's facial expressions. This shows the implementation of HRI (Human-Robot Interaction). In addition to HRI, researchers in [17] used the LBP (Local Binary Patterns) method to extract even low-resolution image features. This makes the facial expression recognition method with LBP can be applied to real-world scenarios, namely videos with low-resolution quality. Besides technology, FER can also be implemented using thermal or infrared photos. Thus, the problem of poor lighting can be resolved so that FER can be helpful in the medical and security world, such as pain or stress detection and night CCTV [18,19]. Lastly, in this research, we try to use preprocessing data method (BiLSTM time-step augmentation) to increase the accuracy of previous work models.[20]

3. Methodology

In this section, we will propose three major phases that consist of dataset gathering, preprocessing data, and proposed models combining proposed CNN and BiLSTM architecture.

3.1. Datasets

The extended Cohn-Kanade, or the CK+ database, is a public facial expression dataset for action units and emotion recognition [21]. This dataset includes both posed and spontaneous expressions. We use The Extended Kohn-Canade (CK+) database, which consists of 981 images obtained from video sequences from a total of 123 different subjects, ranging from 18 to 50 years of age, with various genders and heritage posing seven classes of emotion. The dataset can be accessed at the following link https://www.kaggle.com/datasets/shawon10/ckplus. The seven emotions in the CK+ database are shown in Fig. 2:

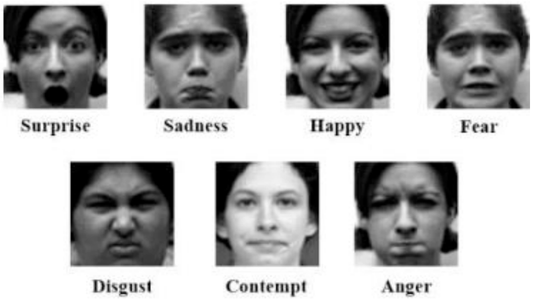


Fig. 2. Seven Emotions in CK+ Database

Table 1 shows how the images in the CK+ datasets are distributed.

Table 1. Distribution Images in CK+ Dataset

Emotion	Frequency
Anger	135
Contempt	54
Disgust	177
Fear	75
Happy	207

Emotion	Frequency
Sadness	84
Surprise	249
Total	**981**

3.2. Preprocessing Data

In this phase, we use min-max normalisation to normalise the images in the CK+ dataset. Min-max normalisation feature a minimum value to be transformed into a 0 and the maximum value into a 1. We define the equation of min-max normalisation as Eq. (1):

$$\frac{X' = X - min(X)}{max(X) - min(X)} \tag{1}$$

Where X' is the normalised data X, min(X) is the minimum value of data X, and max(X) is the maximum value of data X. After the data normalisation, we will also split the data into 10% of data test and 90% of data train. As machine learning, especially deep learning applications grow rapidly, data augmentation techniques have the potential to resolve some of those challenges. Our model will perform better and more accurately if it is fed with a rich, sufficient, and diverse dataset. To achieve the best result for our model, we augment the CK+ dataset using an image data generator from the TensorFlow library.

In this paper, we will build three different models and try to compare them. We will train the CNN, LSTM-CNN, and BiLSTM-CNN models with and without data augmentation. Table 2 shows how different types of preprocessing data are being implemented to suit the built model. A Time step is needed, especially for LSTM and Bidirectional LSTM datasets, because of the time-distributed layer used by the models in the tensor shape.

Table 2. Before and After Augmentation For All Types of Preprocessing Data

Emotions	Types of Preprocessing Data					
	CNN Model		LSTM/Bi-LSTM Model			
	Before Augment	After Augment	Before Augment		After Augment	
	Frequency		Time Step	Frequency	Time Step	Frequency
Anger	135	1451	3	45	3	483
Contempt	54	579	3	18	3	192
Disgust	177	1888	3	59	3	632
Fear	75	797	3	25	3	270
Happy	207	2211	3	69	3	732
Sadness	84	901	3	28	3	297
Surprise	249	2656	3	83	3	888
Total	**981**	**10483**	**3**	**327**	**3**	**3494**

3.3. Proposed CNN Model

In a convolutional neural network, features can be automatically extracted to analyze and classify the images. The Input images are convolved into a feature map or activation map and passed to the new layer. Each feature map's dimensions of data will be reduced to a single neuron at the next layer. Then, every single neuron will connect to another single neuron to produce a fully connected layer. Fig. 3 shows our proposed architecture of CNN (the image is extracted from CK+ database.

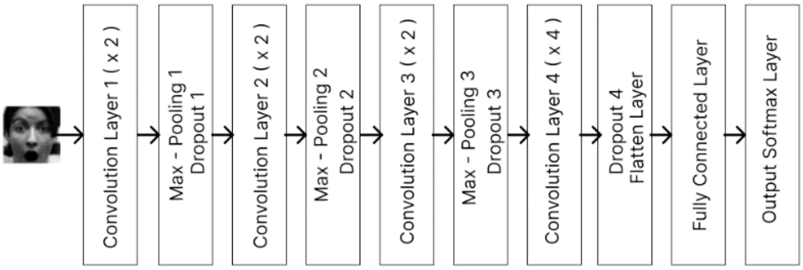


Fig. 3. Proposed CNN Model

The performed CNN will have 13 layers consisting of 8 convolutional layers with 3 x 3 kernel size, 1 stride, and the same padding. The first and second convolutional layers use 64 filters, the third and fourth use 128 filters, the fifth and sixth use 256 filters, and the last two convolutional layers use 512 filters. It aims to extract features like edges, oriented edges, corners, and shapes from the images. Then, we apply 3 max-pooling layers with pool size 2 x 2 after each of the two convolutional layers. The flattened matrix goes through the fully connected layer to produce fully connected layers to classify the seven classes of emotion from the images. We use the ReLu activation function that can be written as Eq. (2):

$$f(x) = max(0, x) \tag{2}$$

The ReLu function f(x) will return 0 if it receives any negative input, but for any positive value x, it returns that value.

3.4. Proposed Model

Our proposed method is the Bidirectional LSTM-CNN model, which combines CNN with BiLSTM to extract the features and process problems involving sequential dependencies from images. Fig. 4 below represents our Bidirectional LSTM-CNN Model:

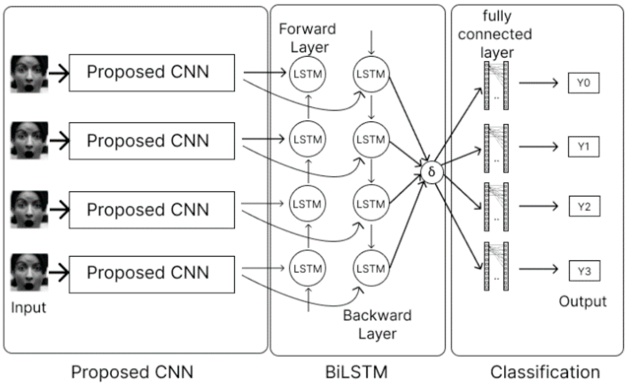


Fig. 4. Proposed CNN Model

Based on Fig. 4, the input features are initially convolved and pooled by the CNN extractor to reduce their dimensions. Also, we use the ELu activation function in the CNN model to avoid the dying ReLu problems. This function tends to converge cost to zero faster and gives better accuracy.

The output from CNN layers is passed to the BiLSTM layer without any loss in accuracy. Then, BiLSTM will feed from both forward and backward directions such that we create a time-distributed CNN within this layer. After that, the output from BiLSTM layers is passed to produce 128 Fully-Connected Layers. This layer will produce the output softmax layer to classify the seven emotion classes.

The training will use 100 epochs and 32 batch sizes with a 0.001 Nadam optimizer. We also define a ModelCheckpoint and CSV_Logger callback in our model to add checkpoints to the trained model and save the model history log. The experiment will use the Keras library in google collab with a capacity of 12.68 GB RAM and 107.72 GB Disk.

4. Result and Discussion

We evaluate our model by the training and validation accuracy and confusion matrix. The accuracy score can be calculated using Eq. (3):

$$Accuracy = \frac{TP}{TP + FP} \tag{3}$$

Where TP is true positive, or the number of facial expressions that are truly predicted, and FP is false positive or the number of facial expressions that are wrongly predicted. In that case, TP+FP means the total amount of facial images in the dataset.

4.1. Model Comparison

Table 3 illustrates the performance of different built models. In overview, the trained model with BiLSTM-CNN architecture performed better than other models before or after data augmentation. The highest accuracy achieved by the BiLSTM-CNN model before augmentation was 81.82%, and after augmentation was 99.43%. The second highest accuracy was the LSTM-CNN model, with 69.70% accuracy before augmentation and significantly increasing to 98.57% after augmentation.

4.2. The Effect of Data Augmentation

Fig. 5 (a) and Fig. 5 (b) show how particularly data augmentation works in our BiLSTM-CNN model. Overfitting often happens between validation and training for both accuracy and loss. Therefore, the data augmentation creates the diversity of the dataset to prevent the overfitting case [22]. We can observe that first, data augmentation will affect

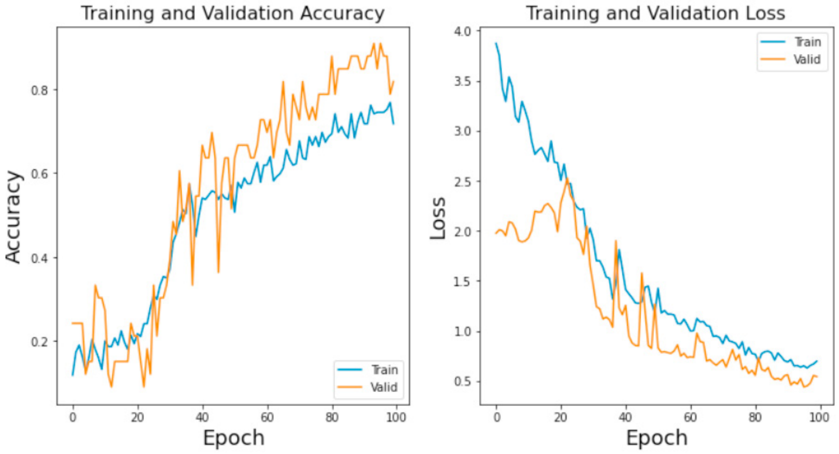


Fig. 5. (a) Loss and Accuracy Before Augmentation.

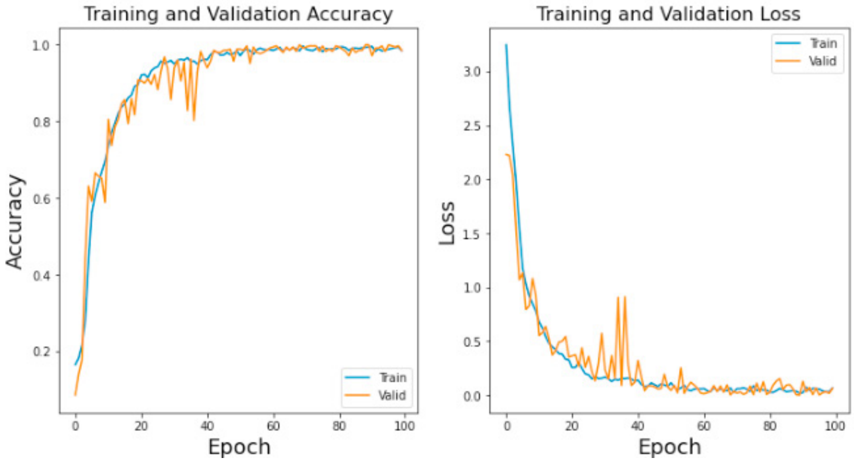


Fig. 5. (b) Loss and Accuracy After Augmentation.

the accuracy of the model, and second, it will solve the overfitting case of the model. Table 3 shows that the accuracy of built models increased significantly after the data augmentation. Also, we can see from Fig. 5 (a) that without data augmentation, the trained model was overfitting because it performed poorly on the validation set. The training and validation for both accuracy and loss were unstable, and the gap between them was also increasing.

Fig. 5 (b) shows how the trained model performed after data augmentation and it fitted properly. We can figure out that the validation was following the training for both accuracy and loss as the number of epochs increased.

Table 3. Comparison of Built Models

Model	Augmentation	Epoch	Batch_Size	Activation	Accuracy (%)
CNN	No	100	32	ReLu	64.65 %
LSTM – CNN	No	100	32	ELu	69.70 %
BiLSTM – CNN	No	100	32	ELu	**81.82 %**
CNN	Yes	100	32	ReLu	80.17 %
LSTM – CNN	Yes	100	32	ELu	98.57 %
BiLSTM – CNN	Yes	100	32	ELu	**99.43 %**

4.3. Comparison to Previous Work

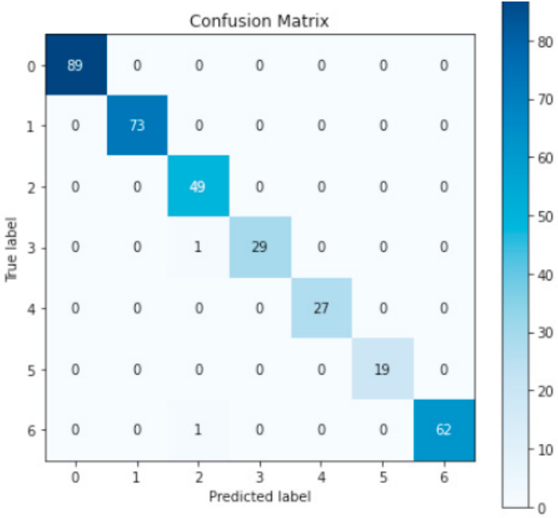
Table 4 illustrates our built model compared with other models from the previous work. We compare our best BiLSTM-CNN model with data augmentation with other methods, namely Inception [23], IB-CNN [24], PPDN [25], and Attentional Convolutional Network [20]. It can be figured out that our best model has reached state of the art and outperforms previous work models in accuracy rate.

Table 4. Comparison of BiLSTM-CNN Model to Previous Work Models

Model	Accuracy (%)
Inception	93.20 %
IB-CNN	95.10 %
PPDN	97.30 %

Model	Accuracy (%)
Attentional Convolutional Network	98 %
BiLSTM - CNN	**99.43 %**

Also, the confusion matrix of our best BiLSTM-CNN model. The confusion matrix is used to summarise and visualise the performance of our model. Each row of the matrix represents emotion in the actual class, while each column represents emotion in the predicted class. Fig. 6 below shows the confusion matrix of our performed model.



	0	1	2	3	4	5	6
0	89	0	0	0	0	0	0
1	0	73	0	0	0	0	0
2	0	0	49	0	0	0	0
3	0	0	1	29	0	0	0
4	0	0	0	0	27	0	0
5	0	0	0	0	0	19	0
6	0	0	1	0	0	0	62

Fig. 6. Confusion Matrix

5. Conclusions

In this paper, we compared built CNN, LSTM-CNN, and BiLSTM-CNN models and tried to augment images in the CK+ dataset to improve the performance of the models. As a result, the BiLSTM-CNN Model with data augmentation achieves the highest accuracy with a value of 99.43%. This implies that data augmentation improves the performance of the trained model and handles the overfitting case. Furthermore, we compare our best model, BiLSTM – CNN, with augmented data from previous work models. The results show that this method achieves a state-of-the-art accuracy rate compared to previous work. For a better result, we will apply the BiLSTM layer to other Computer Vision Deep Learning models such as VGG, Resnet, and SqueezeNet in the future.

References

[1] Pise AA, Alqahtani MA, Verma P, Purushothama K, Karras DA, Prathibha S, et al. Methods for Facial Expression Recognition with Applications in Challenging Situations. Comput Intell Neurosci 2022;2022. https://doi.org/10.1155/2022/9261438.

[2] Mehendale N. Facial emotion recognition using convolutional neural networks (FERC). SN Appl Sci 2020;2:1–8. https://doi.org/10.1007/S42452-020-2234-1/TABLES/3.

[3] Huang Y, Chen F, Lv S, Wang X. Facial Expression Recognition: A Survey. Symmetry 2019, Vol 11, Page 1189 2019;11:1189. https://doi.org/10.3390/SYM11101189.

[4] Jain DK, Shamsolmoali P, Sehdev P. Extended deep neural network for facial emotion recognition. Pattern Recognit Lett 2019;120:69–74. https://doi.org/10.1016/J.PATREC.2019.01.008.

[5] Jee H-K, Jung S-U, Yoo J-H. Liveness Detection for Embedded Face Recognition System 2008. https://doi.org/10.5281/ZENODO.1060812.

[6] An F, Liu Z. Facial expression recognition algorithm based on parameter adaptive initialisation of CNN and LSTM. Visual Computer 2020;36:483–98. https://doi.org/10.1007/S00371-019-01635-4.

[7] Vyas AS, Prajapati HB, Dabhi VK. Survey on Face Expression Recognition using CNN. Undefined 2019:102–6. https://doi.org/10.1109/ICACCS.2019.8728330.

[8] Said Y, Barr M, Ahmed HE. Design of a Face Recognition System based on Convolutional Neural Network (CNN). Engineering, Technology & Applied Science Research 2020;10:5608–12. https://doi.org/10.48084/ETASR.3490.

[9] Shinwari AR, Jalali Balooch A, Alariki AA, Abduljalil Abdulhak S. A Comparative Study of Face Recognition Algorithms under Facial Expression and Illumination. International Conference on Advanced Communication Technology, ICACT 2019;2019-February:390–4. https://doi.org/10.23919/ICACT.2019.8702002.

[10] Yan J, Zheng W, Cui Z, Song P. A joint convolutional bidirectional LSTM framework for facial expression recognition. IEICE Trans Inf Syst 2018;E101D:1217–20. https://doi.org/10.1587/transinf.2017EDL8208.

[11] Pham L, Vu TH, Tran TA. Facial expression recognition using residual masking network. Proceedings - International Conference on Pattern Recognition 2020:4513–9. https://doi.org/10.1109/ICPR48806.2021.9411919.

[12] Shi C, Tan C, Wang L. A Facial Expression Recognition Method Based on a Multibranch Cross-Connection Convolutional Neural Network. IEEE Access 2021;9:39255–74. https://doi.org/10.1109/ACCESS.2021.3063493.

[13] (PDF) Face Trees for Expression Recognition n.d. https://www.researchgate.net/publication/356816965_Face_Trees_for_Expression_Recognition (accessed September 26, 2022).

[14] Ma F, Sun B, Li S. Facial Expression Recognition with Visual Transformers and Attentional Selective Fusion. IEEE Trans Affect Comput 2021. https://doi.org/10.1109/TAFFC.2021.3122146.

[15] Nazeer SA, Omar N, Khalid M. Face recognition system using artificial neural networks approach. Proceedings of ICSCN 2007: International Conference on Signal Processing Communications and Networking 2007:420–5. https://doi.org/10.1109/ICSCN.2007.350774.

[16] Li THS, Kuo PH, Tsai TN, Luan PC. CNN and LSTM Based Facial Expression Analysis Model for a Humanoid Robot. IEEE Access 2019;7:93998–4011. https://doi.org/10.1109/ACCESS.2019.2928364.

[17] Shan C, Gong S, McOwan PW. Facial expression recognition based on Local Binary Patterns: A comprehensive study. Image Vis Comput 2009;27:803–16. https://doi.org/10.1016/J.IMAVIS.2008.08.005.

[18] Kopaczka M, Kolk R, Merhof D. A fully annotated thermal face database and its application for thermal facial expression recognition. I2MTC 2018 - 2018 IEEE International Instrumentation and Measurement Technology Conference: Discovering New Horizons in Instrumentation and Measurement, Proceedings 2018:1–6. https://doi.org/10.1109/I2MTC.2018.8409768.

[19] Kristo M, Ivasic-Kos M. An overview of thermal face recognition methods. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings 2018:1098–103. https://doi.org/10.23919/MIPRO.2018.8400200.

[20] Minaee S, Minaei M, Abdolrashidi A. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. Sensors 2019;21. https://doi.org/10.48550/arxiv.1902.01019.

[21] Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010 2010:94–101. https://doi.org/10.1109/CVPRW.2010.5543262.

[22] Poojary R, Raina R, Mondal AK. Effect of data-augmentation on fine-tuned cnn model performance. IAES International Journal of Artificial Intelligence 2021;10:84–92. https://doi.org/10.11591/IJAI.V10.I1.PP84-92.

[23] Mollahosseini A, Chan D, Mahoor MH. Going deeper in facial expression recognition using deep neural networks. 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016 2016. https://doi.org/10.1109/WACV.2016.7477450.

[24] Han S, Meng Z, KHAN A-S, Tong Y. Incremental Boosting Convolutional Neural Network for Facial Action Unit Recognition. Adv Neural Inf Process Syst 2016;29.

[25] Zhao X, Liang X, Liu L, Li T, Han Y, Vasconcelos N, et al. Peak-Piloted Deep Network for Facial Expression Recognition. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2016;9906 LNCS:425–42. https://doi.org/10.48550/arxiv.1607.06997.