CNN-LSTM Facial Expression Recognition Method Fused with Two-Layer Attention Mechanism

Ye Ming,1,2 Hu Qian,1 and Liu Guangyuan2

1College of Artificial Intelligence, Southwest University, Chongqing, China
2College of Electronic Information Engineering Southwest University, Chongqing, China

Correspondence should be addressed to Ye Ming; zmxym@swu.edu.cn

Received 31 March 2022; Accepted 29 September 2022; Published 13 October 2022

1. Introduction

In the field of facial expression recognition based on deep learning networks, the convolutional neural network (CNN) has been proven to be an important means of extracting local spatial features of images. However, CNN is a typical feed-forward deep network. Its network structure is monotonously connected and the information flows in only one direction, which makes it lack the recognition of contextual timing information and restricts the accuracy of the algorithm.

The recurrent neural network (RNN) has loops because neurons in them not only receive information from other neurons but also receive information from themselves. The input of RNN is sequence data, and a loop unit is executed at each time t. The calculation of the hidden layer at the next time t + 1 depends on the current input data and the hidden state at time t [1]. Since each hidden state saves the node information that is before time t, RNN can establish the global dependency of sequence data. RNN is used in tasks such as speech recognition and language generation.

Computer vision problems involving sequence input can be satisfying solved by RNN. As for multilabel image classification tasks [2], Zhang et al. found that the traditional multilabel image classification method cannot explicitly use the label dependency in the image. It can be seen that on the basis of RNN, CNN-RNN shows a relationship between semantics and image tags. Guo et al. used CNN to gain the discriminative features and used RNN to optimize the classification of coarse and fine labels [3]. Then, the classification performance of traditional hierarchical models is improved by fusing hierarchical information.

Parameters of RNN are adjusted through the back propagation algorithm. However, the network likely experience gradient disappearance as the length of the sequence data input to the network increases. Long short-term memory network (LSTM) introduces gating units to control the accumulation speed of information [1]. This not only endows it with a stronger memory capacity for storing sequence information but also effectively deals with the disappearance of gradients, which is an improvement to RNN.
LSTM has been widely used in image annotation, machine translation, semantic recognition, and other directions [4]. When using neurons as information storage units in neural networks, the more information is stored, the more neurons are needed and the corresponding networks structure are more complex [5]. A large amount of information in the networks often leads to information overload, so the attention mechanism is introduced for this problem. The attention mechanism draws on the basic principles of human brain resource allocation [6]. To obtain richer information from limited resources, more attention is paid to important areas for human attention is limited. Initially, the attention mechanism was proposed in the field of computer vision and later also used in the field of natural language processing [7–9]. Its main purpose is to rationally use computing resources. The method is to allocate more attention to the key parts. This effectively solves the problem of information redundancy and prevents the loss of key information.

Based on the ability of the attention mechanism to focus on key parts and the excellent memory capability of the cyclic neural network, this paper proposes a CNN-LSTM hybrid neural network that integrates a two-layer attention mechanism (ACNN-ALSTM) [10, 11]. The CNN-LSTM network abstracts the local feature information from the facial expression image and establishes the global spatial dependence of the features through the LSTM. At the same time, the two-layer attention mechanism is introduced to mine the information of essential areas to obtain more discriminative expression features, thereby improving the recognition accuracy.

Our contributions are summarized as follows:

1. A two-layer attention mechanism is introduced to propose a novel ACNN-ALSTM model for facial expression recognition. The two-layer attention mechanism can enhance attention to key areas.

2. The LSTM algorithm is introduced to establish the global spatial dependency of facial information. Afterwards, the effect of different network layers and the number of nodes in the hidden layer of the LSTM model on expression recognition are investigated.

3. Comparative experiments are made to verify that the introduction of the two-layer attention mechanism in the CNN-LSTM model improves the classification performance on Fer2013 and CK+ data sets. Comparison with other neural networks that incorporate attention mechanism shows that our work has certain advantages.

2. Related Works

A convolution operator across the spatial domain is added to the network by Yang et al. to extract spatial features to automatically learn spectral spatial features from hyperspectral images [12]. Then, a bidirectional convolutional long and short-term memory network was built using a bidirectional LSTM to fully capture the spectral information. Nagaraju et al. proposed a fully learning-based method for pixel-level segmentation and classification of scene images-2D LSTM network, which achieves good performance on SIFT stream datasets [13], considering the complex spatial dependencies of natural scene image labels.

The model proposed by Lin et al. is composed of a deep LSTM network [14], eight encoders, and eight decoder layers. The attention mechanism is used to connect the bottom layer of the decoder to the top layer of the encoder, which improves parallelism, reduces the training time, and provides a new solution for neural machine translation. Fu et al. found that not all joints can provide information for action analysis in their study of 3D human action recognition [15]. Instead, irrelevant joints usually produce some noise. Attention mechanisms can be employed to focus on the joints that can provide information. Thus, a global context-aware attention LSTM network is proposed to achieve the aim of focusing on the most informative joints in each frame of the skeleton sequence, which shows excellent performance on 3D action recognition datasets.

During recent years, researchers have been trying to optimize the network model using the attention mechanism. For facial expression recognition tasks, Fernandez et al. present a new model, that is end-to-end network architecture with an attention [16]. Li et al. proposed a CNN with attention mechanism (ACNN) [17]. The ACNN can identify the occlusion regions of the facial expression image and focus on the most discriminative regions that are unoccluded.

3. Methodology

In this chapter, the specific composition of the two network architectures we designed is described. For the first one, a layer of attention mechanism is added to the CNN-LSTM network. For the other, a two-layer attention mechanism is introduced.

3.1. CNN-LSTM Facial Expression Recognition Network Model Incorporating an Attention Mechanism. When the RNN is employed for facial expression classification tasks, the results of average sampling according to time steps are used as classification features. However, the average sampling method blurs the focus and cannot expand the advantages of valuable information. The attention mechanism can make the network focus on effective information, so it is employed to selectively process the information input to the classifier to optimize the classification performance of the neural network. The CNN-LSTM facial expression recognition network model incorporating an attention mechanism (CNN-ALSTM) proposed in this section mainly includes four parts: CNN local feature extraction layer, LSTM feature learning layer, global feature attention layer, and classification layer. The CNN-ALSTM model structure diagram is shown in Figure 1.

3.1.1. CNN Local Feature Extraction Layer. This layer is mainly responsible for extracting the local abstract features of the facial expression image and serializing the generated
3.1.2. LSTM Global Feature Learning Layer. The input sequence is $x = \{x_1, x_2, \ldots, x_N\}$, and the input data at moment $n$ is $x_n$. The main work of this layer is roughly divided into four parts. First, the facial expression abstract feature vectors obtained by the CNN feature extraction layer are serialized as the input of the LSTM network. Then, according to the order of the time series, they are sequentially input into the main structure of the recurrent network. At the same time, the LSTM memory unit generates state information combining the current input data and historical information at each moment. Finally, the state information is passed to the next layer feature, which contains historical information at all previous moments.

3.1.3. Global Feature Attention Layer. After the image is extracted by the LSTM network, the hidden state sequence of the LSTM network is defined as $h = \{h_1, h_2, \ldots, h_N\}$, $h_n$ is the hidden state at moment $n$. To select features that are more relevant to the classification task, a query vector $q$ related to the facial expression recognition task is introduced. A scoring function is used to calculate the correlation between the hidden state and the query vector, which means that the higher the correlation between the hidden state and the classification task, the higher the score. For each hidden state, the correlation score can be obtained by the following equation:

$$s(n, q) = h_n^T q.$$  \hfill (1)

$s(n, q)$ is the attention scoring function.

After scoring the relevance of all hidden state tasks, the score is normalized to obtain the attention distribution $a_n$ of the hidden state:

$$a_n = \text{softmax} \left( s(n, q) \right) = \frac{\exp \left( s(n, q) \right)}{\sum_{j=1}^{N} \exp \left( s(j, q) \right)}.$$  \hfill (2)

The attention distribution $a_n$ is the probability representation of the importance of the hidden state $h_n$ in the facial expression recognition task.

The weighted average method is employed to fuse the information of all hidden sequences of the LSTM model, focusing on the most relevant features of the task. The following formula is the weighted feature representation of the sequence of hidden states:

$$\text{att}(h, q) = \sum_{n=1}^{N} a_n h_n.$$  \hfill (3)

Figure 1: CNN-LSTM model the structure of global feature fusing attention mechanism. In CNN local feature extraction layer, C means the convolutional layer and S means the pooling layer. The elements in {} represent the size and the number of the convolution kernels, respectively. The convolutional layers and the pooling layers have step size of 1 and 2, respectively.
3.1.4. Classification Layer. Finally, there is the classification layer. The classification layer is responsible for classifying the weighted features fused by the global feature attention layer.

3.2. CNN-LSTM Facial Expression Recognition Network Model Incorporating a Two-Layer Attention Mechanism. An ACNN-ALSTM network model is proposed in this section. The network mines the local features through CNN [18, 19] and then uses the long-term memory capability of RNN to establish the spatial global dependence of the local features. LSTM is employed to compensate for the network instability when RNN processes long sequences. For the recognition of facial expression features, the proposed model further incorporates the attention mechanism based on the previous section, the process of which is similar to the CNN-ALSTM. The attention distribution is calculated for the hidden state of the LSTM, and then a weighted summation is performed on the hidden state. Afterwards, it is sent to the classifier as an optimized facial expression feature.

3.2.1. CNN Local Feature Extraction Layer. The structure of this layer is exactly the same as the CNN local feature extraction layer in the previous section, so it is not repeated.

3.2.2. Local Feature Attention Layer. The role of this layer is to strengthen the expression ability of essential local features. The query vector \( q \) related to facial expression recognition tasks is introduced. Thus, similar to the previous section, the correlation between each local feature \( \text{local}_n \) and the query vector can be calculated through the scoring function. Local features with high scores are more essential:

\[
s(\text{local}_n, q) = \text{local}_n^T q, \quad (4)
\]

\( S(\text{local}_n, q) \) is the attention scoring function. After completing the task relevance scoring, the score is normalized to obtain the attention distribution \( L_a_n \) of the local features as shown in the following formula:

\[
L_{a_n} = \text{softmax} \left( s(\text{local}_n, q) \right). \quad (5)
\]

The attention distribution \( L_{a_n} \) is the probability representation of the essential of the local feature \( \text{local}_n \). Then, weight each feature as the input of LSTM:

\[
x_n = L_{a_n} \cdot \text{local}_n. \quad (6)
\]

\( x_n \) is the weighted local feature.

3.2.3. LSTM Global Feature Learning Layer. The process at this layer is similar to that described in the previous section. The LSTM algorithm employs gating units and strong memory ability to establish the global spatial dependence of facial information.

3.2.4. Global Feature Attention Layer. The hidden state output by LSTM is employed as the input of the global feature attention layer. The query vector \( q \) of the attention layer is defined as a learnable weight vector whose dimension is consistent with the number of hidden layer nodes of the LSTM network.
This layer implements the weighted fusion of the output features of the LSTM using the attention mechanism in two steps. First, the attention distribution \( a = \{a_1, a_2, \ldots, a_N\} \) for the hidden state \( h = \{h_1, h_2, \ldots, h_N\} \) of the LSTM network is calculated. Second, the weighted average result \( att \) of the hidden state is calculated according to the attention distribution.

The attention mechanism can increase the weight of the features that are most relevant to facial expression recognition tasks and achieve its dominant role in the feature fusion process, thereby improving the classification performance.

3.2.5. Classification Layer. The classification layer is responsible for classifying the weighted features fused by the global feature attention layer. First, to reduce the dimensionality of the feature to the number of categories, the full global feature attention layer. First, to reduce the dimensionality of the feature to the number of categories, the full global feature attention layer. Finally, the weighted average result \( att \) of the hidden state is calculated according to the attention distribution.

The attention mechanism can increase the weight of the features that are most relevant to facial expression recognition tasks and achieve its dominant role in the feature fusion process, thereby improving the classification performance.

Finally, the loss of the entire model is calculated by cross entropy, and the calculation method is shown as follows:

\[
L(y, S) = -y \log(S).
\]  

\( y \) refers to the probability distribution in the real category, and \( S \) represents the probability distribution of the categories predicted by the model in this section.

4. Experiments

The data sets, simulations implementation, training details, and expression recognition results are described carefully in this part. As for the analysis of the experimental results, the effect of LSTM parameters on the facial expression features extracted by the model is experimentally studied. Besides, the performance of the methods proposed is confirmed by the expression recognition result.

4.1. Datasets. Two data sets are employed, namely, CK+ and Fer2013 data sets. Both data sets contain facial images of 7 emotions and they will be described in detail.

4.1.1. Fer2013 Data Set [20]. The Fer2013 data set is mainly from the well-known data science competition platform Kaggle. The data is obtained by searching for image keywords through the Google search engine, including 35,887 gray-scale images with a resolution of 48 \( \times \) 48 pixels. Table 1 shows the details of various expressions in the Fer2013 data set. The data set is rich and diverse since the images are crawled from the network, but the data contains a lot of noise, which is a challenge for facial expression recognition classification algorithms.

4.1.2. CK+ Data Set. The CK+ (Extended Cohn Kanade) data set was established in 2010 by the Patrick Lucey team and the Zara Ambadar team [21]. The database is a close-up of faces taken by 123 subjects according to certain expression requirements. There are 593 facial image data in the database, 327 of which include seven kinds of facial emotions. Since there is less data with expression tags in this data set, data enhancement processing is performed on 327 images with emotions [22]. First, face recognition is performed on the image to trim the invalid background. Then, all images are converted to 48 \( \times \) 48 resolution. Furthermore, the images are flipped and rotated to increase the amount of data to about 3 times the original. At the same time, the brightness and saturation of the images are adjusted. The data distribution of the processed CK+ data set is shown in Table 2.

4.2. Implementation and Training Details. Two kinds of image classification experiments are designed to verify the effects of the two deep learning network models proposed in the paper on facial expression recognition. These experiments are implemented on the Linux16.04 64 bit operating system, using python3.6 and the Google framework tensorflow1.12.0-gpu for programming and implementation, and calling OpenCV for data enhancement. NVIDIA GeForce GTX1080ti graphics card with 16G video memory is employed. The FER2013 and CK+ data sets are divided into

### Table 1: Fer2013 data set.

| Expression | Total | Anger | Disgust | Fear | Glee | Dump | Surprise | Neutral |
|------------|-------|-------|--------|------|------|------|---------|---------|
| Tag        | —     | 0     | 1      | 2    | 3    | 4    | 5       | 6       |
| Quantity   | 35887 | 4953  | 547    | 5121 | 8989 | 6077 | 4002    | 6198    |

### Table 2: Data distribution after data enhancement for CK+ (extended Cohn Kanade) data set.

| Expression | Total | Anger | Disgust | Fear | Glee | Dump | Surprise | Despise |
|------------|-------|-------|--------|------|------|------|---------|---------|
| Quantity   | 981   | 135   | 177    | 75   | 207  | 84   | 249     | 54      |
training set, validation set, and test set according to the ratio of 8:2. The iteration is 5000 and print the training loss and error every 100 rounds. The loss and error of the test are printed every 400 rounds. Face images of the input model are rescaled to $48 \times 48$ pixels. The learning rate of the model is 0.001, and Adam optimizer and softmax regression are employed for calculation and classification [23, 24]. Cross entropy is used to calculate the loss function, and the batch size is 30.

It is found that the feature abstraction degree of the LSTM model of the image is related to the number of nodes in the hidden layer and the depth of the hidden layer. Therefore, the influence of the number of nodes in the hidden layer and the depth of the hidden layer on the effect of facial expression recognition is shown in Table 3.

4.3. Expression Recognition Results. We conducted two types of experiments. First of all, the experiment studied the influence of LSTM parameters on facial expression features extracted by the model. Then, a comparison about various variants of ACNN-ALSTM and related work is executed.

4.3.1. The Influence of LSTM Model Parameters on Facial Expression Recognition. The paper considers two essential factors that affect the feature extraction of LSTM: the number of network layers and the number of hidden layer nodes. The classification results of differently constructed LSTM networks on the Fer2013 and the processed CK + data sets are shown in Figure 3.

According to the experimental results, the classification accuracy rate continues to increase as the number of nodes increases. At the same time, the abstract features learned from the image data of the LSTM network are more discriminative. However, the increase in the number of hidden layer nodes has an upper limit on the effect of increasing the recognition rate. From Figure 3, it can be found that the effect is the same when the hidden layer nodes are set to 512 and 1024. Besides, increasing the number of network layers can improve the facial expression recognition effect of the model. When the LSTM model is a two-layer network and the number of hidden layer nodes is 512-512, the classification accuracy of the model on the Fer2013 and CK + datasets is relatively high.

4.3.2. Comparison of Facial Expression Recognition Results of ACNN-ALSTM. CNN and CNN-LSTM models. The CNN-LSTM hybrid neural network is proposed by combining the advantages of CNN and LSTM to mine image features. The
CNN and the CNN-LSTM models are demonstrated in Figures 4(a) and 4(b), and the two figures show the classification accuracy on Fer2013 and the processed CK+ dataset after every 100 trainings, respectively. On the Fer2013 training set, the classification accuracy of the CNN-LSTM hybrid model surpassed the CNN model after 1900 iterations. On the CK+ training set, the training classification accuracy of the CNN-LSTM hybrid model is always better than the CNN model. It can be seen that the CNN-LSTM hybrid model can effectively recognize facial expressions.

CNN-LSTM model and its various variants. These models are trained on the Fer2013 and the processed CK+ training sets. Table 4 shows the average test results obtained through multiple experiments on the test set. In the table, ACNN-LSTM represents that the local feature is fused with the attention mechanism. The CNN-ALSTM represents that the global feature is fused with the attention mechanism. The ACNN-ALSTM model represents that the local feature and the global feature are both fused with the attention mechanism.

Compared with that only the local feature or that only global feature is fused with the attention mechanism, the effect of both local feature and global feature are fused with the attention mechanism is better. At the same time, it can be found that the effect of fusing the attention mechanism in the global feature is better than that in the local feature. The above-given research results show that adding an attention mechanism to the CNN-LSTM model can highlight essential features and improve the classification effect.

ACNN-ALSTM and related work. Although some results cannot be directly compared due to different experimental setups and different preprocessing methods, it is demonstrated in Table 5 that the proposed method can yield a feasible and promising recognition rate for static facial images.

5. Discussion

It is obvious that the efficiency and accuracy of recognition can be promoted by making sufficient use of key parts of facial expression. However, the utilization of the key parts of the existing algorithm can be further improved. Thus, the CNN-ALSTM and the ACNN-ALSTM incorporating an attention mechanism are proposed in this paper. Simulation results indicate that the proposed algorithm can further improve the recognition rate compared with other researchers’ work, although both their algorithms and ours incorporating attention mechanisms. At the same time, this work demonstrates the effectiveness of attentional mechanisms for exploiting key information about facial expressions.

6. Conclusion

In this work, a novel ACNN-ALSTM model for facial expression recognition that incorporates a two-layer attention method to the CNN-LSTM model can highlight essential features and improve the classification effect.

Table 4: Comparison of classification accuracy of four models.

| Method     | Fer2013 (%) | CK+ (%) |
|------------|-------------|---------|
| CNN-LSTM   | 74.1        | 95.3    |
| ACNN-LSTM  | 75.5        | 97.7    |
| CNN-ALSTM  | 77.4        | 99.1    |
| ACNN-ALSTM | **78.2**    | **99.7**|

Table 5: Comparison of classification accuracy with related work.

| Method      | Fer2013 (%) | CK+ (%) |
|-------------|-------------|---------|
| pACNN [17]  | 74.3        | 97.0    |
| FERAtt [16] | 74.8        | 97.5    |
| ACNN-ALSTM  | **78.2**    | **99.7**|
mechanism is proposed. Comparative experimental results indicate that the introduction of the two-layer attention mechanism improves the classification performance of the system more significantly than that of the CNN-LSTM model and its variants. We also investigate the effect of the different network layers and the number of nodes in the hidden layers of the LSTM model on expression recognition. The results indicate that appropriately increasing the network depth and the number of hidden layer nodes is beneficial to improve the recognition accuracy of the LSTM model. In future work, we want to train the network for extreme conditions such as dark, light, and occlusion [25, 26].

Data Availability

https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant nos. 61472330 and 31571111).

References

[1] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," Physica D: Nonlinear Phenomena, vol. 404, Article ID 132306, 2020.
[2] L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Z. Li, “Face detection based on multi-block lbp representation,” in International Conference on Biometrics, pp. 11–18, Springer, Berlin, Heidelberg, 2007.
[3] Z. Guo, L. Zhang, and D. Zhang, "Rotation invariant texture classification using LBP variance (LBVP) with global matching," Pattern Recognition, vol. 43, no. 3, pp. 706–719, 2010.
[4] Y. Yu, X. S. Si, C. H. Hu, and J. X. Zhang, “A Review of recurrent neural networks: LSTM Cells and network architectures,” Neural Computation, vol. 31, no. 7, pp. 1235–1270, 2019.
[5] M. Cai and J. Liu, "Maxout neurons for deep convolutional and LSTM neural networks in speech recognition," Speech Communication, vol. 77, pp. 53–64, 2016.
[6] T. Moore and M. Zirnsak, "Neural mechanisms of selective Visual attention," Annual Review of Psychology, vol. 68, no. 1, pp. 47–72, 2017.
[7] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level CNN: Salience-aware 3-D CNN with LSTM for video action recognition," IEEE Signal Processing Letters, vol. 24, no. 4, pp. 510–514, 2017.
[8] F. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "EleAtt-RNN: adding Attentiveness to neurons in recurrent neural networks," IEEE Transactions on Image Processing, vol. 29, pp. 1061–1073, 2020.
[9] R. Hang, Z. Li, Q. Liu, P. Ghamisi, and S. S. Bhattacharyya, "Hyperspectral image classification with attention-Aided CNNs," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 3, pp. 2281–2293, 2021.
[10] T. Li, M. Hua, and X. Wu, "A hybrid CNN-LSTM model for forecasting Particulate Matter (PM2.5)," IEEE Access, vol. 8, pp. 26933–26940, 2020.
[11] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomedical Signal Processing and Control, vol. 47, pp. 312–323, 2019.
[12] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 1, pp. 131–137, 2004.
[13] C. Nagaraju, D. Sharadamani, C. Maheswari, and D. Vishnu Vardhan, “Evaluation of LBP-based facial emotions recognition techniques to make consistent decisions,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 29, no. 06, Article ID 1556008, 2015.
[14] K. Lin, Y. Xu, and Z. Zhong, “Using kernel discriminant analysis and 2Dgaabor local features fusion for face recognition,” International Journal of Digital Content Technology and its Applications, vol. 4, no. 8, pp. 232–241, 2010.
[15] S. Fu, X. Kuai, and G. Yang, “Facial expression recognition by independent log-Gabor component analysis,” in International Symposium on Neural Networks, pp. 305–312, Springer, 2011.
[16] P. D. M. Fernandez, F. A. G. Pena, T. I. Ren, and A. Cunha, “Feratt: facial expression recognition with attention net,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, June 2019.
[17] Y. Li, J. Zeng, S. Shan, and X. Chen, “Occulsion aware facial expression recognition using cnn with attention mechanism,” IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2439–2450, 2019.
[18] H. Tang, B. Xiao, W. Li, and G. Wang, “Pixel convolutional neural network for multi-focus image fusion,” Information Sciences, vol. 433, pp. 125–141, 2018.
[19] Y. Wei, W. Xia, M. Lin et al., “HCP: a Flexible CNN framework for multi-label image classification,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 9, pp. 1901–1907, 2016.
[20] H. D. Nguyen, S. Yeom, G.-S. Lee, H.-J. Yang, I.-S. Na, and S.-H. Kim, “Facial emotion recognition using an Ensemble of multi-level convolutional neural networks,” International Journal of Pattern Recognition and Artificial Intelligence, vol. 33, no. 11, Article ID 1940015, 2019.
[21] 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 Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops CVPRW 2010, pp. 94–101, IEEE Computer Society, San Francisco, CA, United states, June 2010.
[22] 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 Recognition, vol. 61, pp. 610–628, 2017.
[23] S. Bera and V. K. Shrivastava, “Analysis of various optimizers on deep convolutional neural network model in the
application of hyperspectral remote sensing image classification," *International Journal of Remote Sensing*, vol. 41, no. 7, pp. 2664–2683, 2020.

[24] K. K. Chandriah and R. V. Naraganahalli, "RNN/LSTM with modified Adam optimizer in deep learning approach for automobile spare parts demand forecasting," *Multimedia Tools and Applications*, vol. 80, no. 17, pp. 26145–26159, 2021.

[25] G. Batchuluun, D. T. Nguyen, T. D. Pham, C. Park, and K. R. Park, "Action recognition from Thermal videos," *IEEE Access*, vol. 7, pp. 103893–103917, 2019.

[26] Q. Wang, C. Yuan, J. Wang, and W. Zeng, "Learning attentional recurrent neural network for Visual Tracking," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 930–942, 2019.