Computers, Materials & Continua
DOI:10.32604/cmc.2021.016832
Article

Tech Science Press

Emotion Analysis: Bimodal Fusion of Facial Expressions and EEG

Huiping Jiang[1,*], Rui Jiao[1], Demeng Wu[1] and Wenbo Wu[2]

[1]Brain Cognitive Computing Lab, School of Information Engineering, Minzu University of China, Beijing, 100081, China
[2]Case Western Reserve University, USA
[*]Corresponding Author: Huiping Jiang. Email: jianghp@muc.edu.cn
Received: 13 January 2021; Accepted: 24 February 2021

Abstract: With the rapid development of deep learning and artificial intelligence, affective computing, as a branch field, has attracted increasing research attention. Human emotions are diverse and are directly expressed via non-physiological indicators, such as electroencephalogram (EEG) signals. However, whether emotion-based or EEG-based, these remain single-modes of emotion recognition. Multi-mode fusion emotion recognition can improve accuracy by utilizing feature diversity and correlation. Therefore, three different models have been established: the single-mode-based EEG-long and short-term memory (LSTM) model, the Facial-LSTM model based on facial expressions processing EEG data, and the multi-mode LSTM-convolutional neural network (CNN) model that combines expressions and EEG. Their average classification accuracy was 86.48%, 89.42%, and 93.13%, respectively. Compared with the EEG-LSTM model, the Facial-LSTM model improved by about 3%. This indicated that the expression mode helped eliminate EEG signals that contained few or no emotional features, enhancing emotion recognition accuracy. Compared with the Facial-LSTM model, the classification accuracy of the LSTM-CNN model improved by 3.7%, showing that the addition of facial expressions affected the EEG features to a certain extent. Therefore, using various modal features for emotion recognition conforms to human emotional expression. Furthermore, it improves feature diversity to facilitate further emotion recognition research.

Keywords: Single-mode and multi-mode; expressions and EEG; deep learning; LSTM

1 Introduction

Emotion can be described as a sudden response to external or internal events and occurs instinctively. Emotions have always played an essential role in human life, work, and decision-making. With the development of deep learning and artificial intelligence, the prospect of emotion recognition in the field of human-computer interaction is broader. Emotion recognition can be achieved using facial expressions, tone of voice, motion, and physiological signals [1,2].

Facial expressions are the most direct form of human emotional manifestation and are reflected in the mouth, cheeks, eyes, and other facial features. Therefore, most researchers use facial expressions as a starting point to analyze emotional changes [3]. Lu et al. [4] used principal component analysis (PCA) to reduce dimensionality with a support vector machine (SVM) as the classifier, producing a classification result of 78.37%. Qin et al. [5] proposed a method that combined the Gabor wavelet transform and CNN, resulting in a 96.81% accuracy on the CK + data set. Rajan et al. [6] combined CNN with LSTM units for real-time facial expression recognition (FER), effectively utilizing time and space features.

Although facial expressions can directly reflect personal emotions and are readily obtainable, they are easy to conceal, hide, or provide single data. Therefore, facial expressions sometimes do not reliably reflect true emotions, a common defect of non-physiological signals. Consequently, researchers examine physiological signals instead. Neurophysiologists and psychologists have found that the physiologically manifested EEG signals are closely related to most emotions [7]. Zhang et al. [8] combined wavelets and CNN to classify emotions, with the best effect reaching 88%. Zhang et al. [9] proposed a method using CNN for the emotion recognition of EEG signals. This showed that CNN could autonomously extract features from signals. Alhagry et al. [10] used LSTM to learn and classify EEG signal features, obtaining 87.99% classification accuracy in the Database for Emotion Analysis using Physiological Signals (DEAP) data set.

Although the original EEG signals can provide useful emotional information, solely relying on them for emotion recognition is challenging due to weak signal strength. Whether they are utilized to recognize facial expression modalities or EEG, the expression forms of these signals are relatively straightforward [11]. Expression and EEG signals have been extensively examined in a non-physiological and physiological context and can be effectively combined for multi-modal emotion recognition. Therefore, this synergistic relationship allows the complementary information to improve the objectivity and accuracy of emotion recognition [12]. Shu et al. [13] proposed a fusion strategy based on the decision matrix in the study of multi-modality to improve the accuracy of the system. Huang [14] proposed two decision-level fusion methods for EEG and facial expression detection. The accuracy rates are 81.25% and 82.75%, respectively.

Combining facial expressions and EEG information for emotion recognition compensates for their shortcomings as single data sources [15,16]. This paper realizes emotion recognition via a modal fusion of facial expressions and EEG data. Since decision fusion does not make fair use of the correlation between different modalities, the method used in this paper involves feature-level fusion. The work content is as follows:

(a) This paper establishes a model-facial expression recognition system for multi-modal emotion recognition.
(b) It is expected to add expression information for single-modal EEG emotion recognition. Compared with the original single-modal EEG emotion recognition results, the multi-modal method is found to be superior.
(c) This article proposes two different ways of combining facial expressions and EEG information for sentiment analysis. The Facial-LSTM model refers to EEG data processing during the first and last facial expression change frames by the model-facial expression recognition system. The LSTM-CNN model feeds the preprocessed facial expressions and EEG data into the LSTM for feature extraction. The output features are then fused and sent to the CNN for classification.

2 Related Work

2.1 LSTM

A challenge is presented by the fact that the recurrent neural network (RNN) is incapable of long-term memory due to gradient disappearance or gradient explosion. Schimidhuber et al. [17] improved the traditional RNN and proposed the LSTM to solve this problem. The LSTM introduces an additional memory unit, C. This is a self-connecting unit that can store long-term signals and help LSTM encode distant historical information. Fig. 1 shows the LSTM memory unit, where subscripts t and $t-1$ represent the current and previous moments.

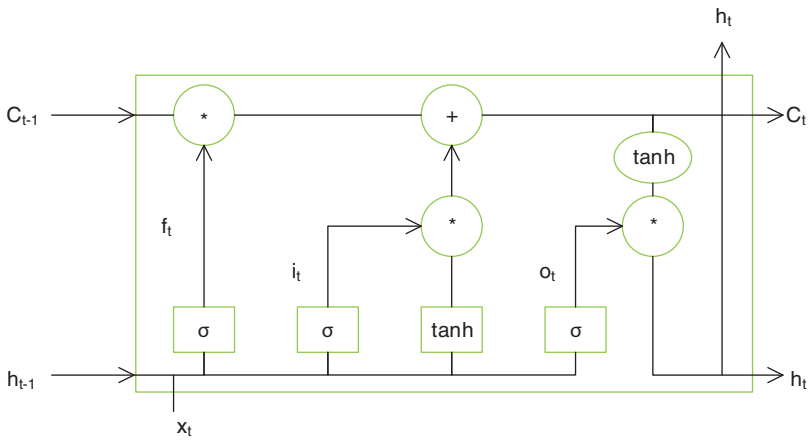

Figure 1: An LSTM memory unit

The calculation process of the LSTM unit occurs as follows:

Discarding the gate: the h_{t-1} state before the new x_t input determines the portion of C information that can be discarded. The f_t and C_{t-1} gates are discarded to calculate and remove part of the information. The σ operator represents the sigmoid operation, 0 for discard, and 1 for save, which are used to determine the parameter change in C_{t-1}. The calculation formula is shown in Eq. (1):

$$f_t = \sigma\left(\mathrm{w}_{fx}x_t + w_{fh}h_{t-1} + \mathrm{w}_{fc}C_{t-1} + b_f\right) \tag{1}$$

Here, w_{fx}, w_{fh}, and w_{fc} represent the weight of the forgetting door, the memory door of the LSTM unit at the previous moment, and the memory unit of the forgetting door at the previous moment, respectively, while b denotes the bias quantity.

The input gate: The h_{t-1} state before the new x_t input determines the information saved by C. The calculation formula for the i_t input gate is shown in Eq. (2):

$$i_t = \sigma\left(\mathrm{w}_{ih}h_{t-1} + w_{xi}x_t + \mathrm{w}_{ci}C_{t-1} + b_i\right) \tag{2}$$

Here, i_t signifies the control parameter of the $\widetilde{C}_t$ coefficient when new information is added, which is used to update C. w_{xi}, w_{ih} and w_{ci} represent the weight of the input gate, the input gate of the LSTM unit at the previous moment, and the input gate memory unit at the previous moment, respectively, while b denotes the bias quantity.

Updating the control parameters: According to the old C_{t-1} control parameters, a new generation of $\widetilde{C}_t$ control parameters are generated the moment the final control parameters, as shown in Eq. (3):

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{3}$$

In the above equation, the status update of the memory unit depends on its own C_{t-1} state. The current candidate memory unit value, $\widetilde{C}_t$,i is adjusted by the input and discarding gates.

The output gate: The new LSTM output is generated according to the C_t control parameter, as shown in Eq. (4):

$$o_t = \sigma\left(\mathrm{w}_{xo} x_t + w_{ho} h_{t-1} + \mathrm{w}_{co} C_{t-1} + b_o\right) \tag{4}$$

Here, o_t represents the state value of the control memory unit. w_{xo}, w_{ho}, and w_{co} correspond to the weight of the corresponding output gate, the output gate of the LSTM unit at the previous moment, and the output gate at the previous moment, respectively, while b denotes the bias quantity, as shown in Eq. (5):

$$h_t = o_t * \tanh\left(C_t\right) \tag{5}$$

By introducing a gating design, LSTM can effectively eliminate the RNN gradient disappearance, allowing the RNN model to be effectively applied to the long-distance sequence information.

2.2 Facial Expression Recognition System (Model-Facial)

Emotional changes in the EEG signals do not occur continuously. Therefore, there are periods when the EEG signals do not contain enough emotional information. Facial expression data is useful for obtaining emotional information, therefore, establishing a facial expression recognition system. The model-facial is divided into two parts: the training module and the recognition module [18].

The training module process is shown in Fig. 2.

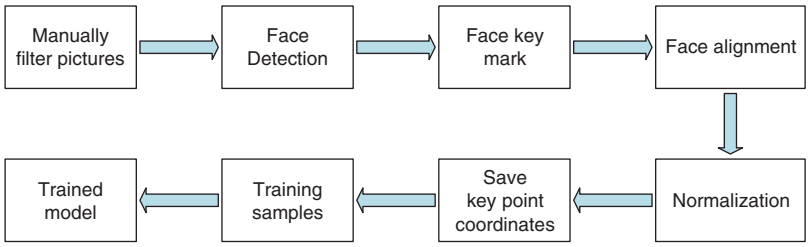

Figure 2: A flow chart of the Model-Facial training module

The recognition module process is shown in Fig. 3.

The videos collected during the experiment were segmented into pictures by frame. Here, 100 images of facial expressions depicting obvious calm, negative, and positive emotions were selected for each subject. The experiment relied on the critical facial point detection model of the open-source Dlib library. The detection results involving 68 key facial points are shown in Fig. 4. Following image normalization and facial alignment, the horizontal and vertical coordinates of

the 68 points were saved. A total of 136 values were entered into a text file, representing the extracted features [19].

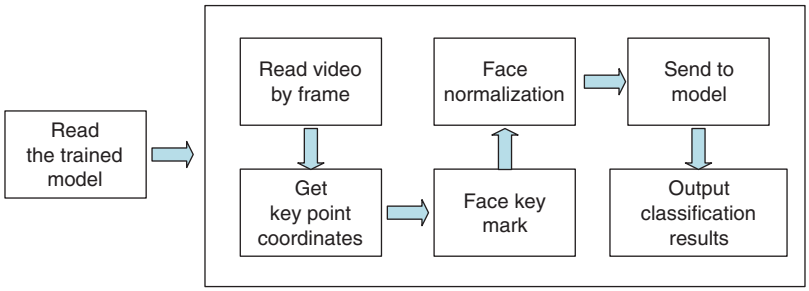

Figure 3: A flow chart of the Model-Facial recognition module

Figure 4: An example of the 68 facial keypoint detection results

SVM was used during the experiment to classify the extracted features. After model training, the expression videos of the subjects were read according to the frame, classifying each frame. All the participants were fully aware of the purpose of the study, which was approved by the Local Ethics Committee (Minzu University of China, Beijing, ECMUC2019008CO).

2.3 Multi-Modal Fusion

In the field of emotion recognition, emotions exhibit various modes. For example, gestures, expressions, words, or other physiological signals (EEG and Electrocardiograph) can express the emotions of an individual. Although these modes can reflect independent feelings, humans generally express multiple emotions simultaneously during interaction with others [20]. Multi-modality can provide more comprehensive and accurate information, enhancing the reliability and fault tolerance of the emotion recognition system. Unlike single-mode emotion recognition, the multi-modal form involves obtaining single-modal expression while better utilizing the correlation between the various modes to combine them. This is known as modal fusion.

Furthermore, multi-modal fusion methods can be divided into signal-level fusion, feature level fusion, and decision level fusion according to the processing of different modal signals.

As the name suggests, signal-level fusion directly combines and processes the originally collected signals and then performs feature extraction and recognition [21]. This fusion method retains the original signal, ensuring that the accuracy is high. Moreover, its low anti-interference ability can be ascribed to a substantial amount of data collected for a long time.

Feature layer fusion refers to the fusion and classification of features extracted from single modes. This technique takes advantage of the correlation between features to a greater extent. However, feature layer fusion requires an exceedingly high synchronization of the collected data.

Decision level fusion refers to extracting features from single modes, classifying them before fusion judgment. The most significant advantage of the decision layer is that it simplifies merging the decisions acquired from each pattern. Flexibility is increased since each mode can learn its characteristics using the most appropriate classification model. However, decision level fusion does not take advantage of the correlation between modal characteristics.

Given the differences in the modal characteristics of the expression and EEG signals, utilizing the signal level for fusion is challenging [22]. Decision level fusion does not consider the correlation between the two parts. Therefore, this paper selected the fusion of feature layer to realize the bi-modal fusion of expression and EEG data, as shown in Fig. 5.

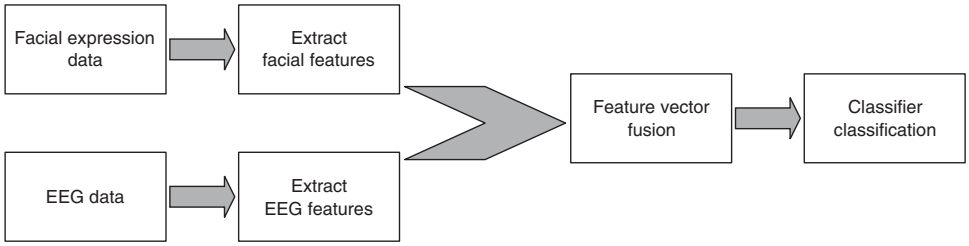

Figure 5: Feature level fusion

3 Experimental

3.1 Stimulus Materials

The collection of the original data is mainly divided into two parts: emotion induction and data collection. The emotional installation refers to dynamically stimulating the subject to produce the target emotion. Data collection includes the acquisition of EEG signals and expressions.

Dynamic stimulation materials combined with visual and auditory stimulation characteristics have a better effect on emotional induction. Therefore, 12 videos with positive and negative colors were initially screened. Non-participants then completed questionnaires that screened out the video material used in the formal study. Pleasure and surprise were denoted as positive emotions according to the two-dimensional emotion model and the perspective of Fred Ekman et al., while several other emotions were designated as negative. Ultimately, six videos portraying high emotional arousal and induction intensity were identified among the 12 videos. Three of these were positive, and three were negative.

The subjects selected for the experiment were all students aged between 18 and 25, right-handed, in good physical condition, fully rested, and free of brain diseases or mental problems.

Before the formal experiment, the subjects were required to read the instructions on display carefully, ensuring they fully understood the experimental process. During the experiment, the

EEG data and the corresponding facial information of the experimenter were also obtained and saved.

3.2 Collection and Pretreatment

The facial expressions were primarily collected using a Logitech Carl Zeiss Tessar camera in conjunction with EV video recording. The resolution was $1920 * 1080$, and the video acquisition frequency was 13 fps. The EEG acquisition and recording were performed using the NeuroScan system platform. E-prime was employed to design and present the stimulus materials to the subjects, triggering emotional changes. During the experiment, a 64-electrode cap was used to collect EEG information, which was expanded using an amplifier, and recorded on a computer equipped with scanning software.

The EEG signal was so weak that it was highly susceptible to the internal and external environment during measurement. This rendered the collected signal unreliable due to disturbance by considerable electrical activity not originating from the brain, known as artifacts. The artifacts were commonly initiated by electrooculograms, electrocardiograms, electromyograms, and electrode motion. Scan 4.5 was used to complete EEG preprocessing, including removing the bad areas, ophthalmological artifacts, and other artifacts, as well as digital filtering. Preprocessing primarily eliminated the noise component of the EEG signal, preparing it for feature analysis or extraction of the emotional elements.

The original facial expression data was presented in the form of a video, which included the emotional changes detected in the subjects. Therefore, the first step in preprocessing the facial expression data was to cut the tape into frames. The second step involved frame selection. There is no expression data in the before and after video, so exclude some frames in this part. The third step involved facial detection. Factors other than human faces were removed to improve facial feature extraction and noise reduction.

3.3 Training

The emotion classification model based on LSTM consisted of four layers, including the input layer, the LSTM layer, the full connection layer, and the output layer. The LSTM layer extracted the relevant features from the EEG input sequence, also known as the time-domain information. The full connection layer integrated the components of the LSTM layer, obtaining the desired classification results.

While establishing the LSTM layer, it was necessary to select the appropriate number of layers and determine the number of hidden nodes in each. Generally, the presence of many neurons causes overfitting during training, while a small number of neurons may result in underfitting. This necessitated designing as few hidden layer nodes and LSTM layers as possible under the premise of meeting the accuracy requirements. The experiment involving the single- and multi-layer LSTM structure revealed that the latter exhibited a higher classification effect. Finally, it was determined that each layer of the model was formed by 32 hidden nodes in series, consisting of four layers, as shown in Fig. 6.

The Adam algorithm was adopted for parameter optimization, while the learning rate of the model was set to 0.005. The Dropout method was used during neural network training to avoid the overfitting phenomenon, while the parameter value was set to 0.5. The batch processing technology was used during the training process to determine the batch size of 64 training samples. Google's TensorFlow framework was employed to implement the network model.

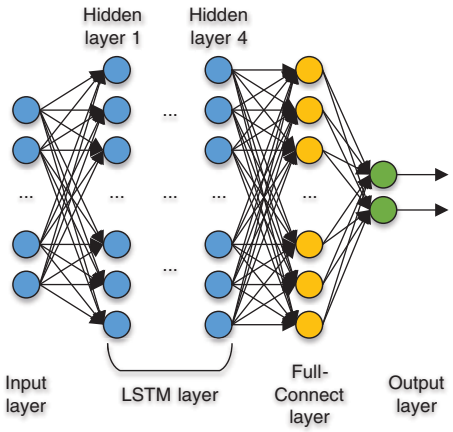

Figure 6: LSTM structure

The specific parameter settings of the LSTM emotion classification model are shown in Tab. 1.

Table 1: Parameter settings of LSTM model

Name	Parameter
Learning rate	0.005
Input dimension	64×10
Output category	2
Batch	64
Dropout	0.5
Hidden node number	32
LSTM layer number	4

3.3.1 EEG-LSTM

The EEG-LSTM model represents the single-mode of EEG emotion recognition. After preprocessing, the EEG signal is sent to the LSTM for classification.

The dataset in this section is the complete EEG data of eight subjects. After simple preprocessing, the EEG data of each subject were divided at intervals of 10 ms, while each EEG data set collected by the 64 conducting pole caps had a $64*10$ matrix dimension. According to the LSTM principle, the matrix columns represent the data read in one step, while the rows represent the time step. This way of intercepting EEG data produced a large enough amount of information. The ratio of the train set to the test set was about 3:1. Tab. 2 shows the classification effect of 8-bit subjects.

3.3.2 Facial-LSTM

The Facial-LSTM Model refers to the model-facial expression recognition system used to tailor the EEG data. The cropped EEG information is sent to the LSTM emotion classification

model for emotion recognition. Since the expression of a subject does not remain static in a calm state, the movement, relaxation, and twitching of the facial muscles in a natural state also cause expressional changes. Therefore, it was proposed to set the first frame where the facial expression of the subject changed for more than five consecutive frames as the starting keyframe. Similarly, the last frame in which the facial expression of the subject changed for five successive frames or more was set as the end keyframe.

Tab. 3 shows the first and last keyframes of specific subjects.

Table 2: Classification results of the EEG-LSTM model

Experimenter	Accuracy (%)
Subject1	84.72
Subject2	84.57
Subject3	96.04
Subject4	84.57
Subject5	90.36
Subject6	93.02
Subject7	81.91
Subject8	76.64
AVG	86.48

Table 3: Keyframes of specific subjects

	Start frame	Start time (ms)	End frame	End time (ms)
Video1	885	68076	3098	238307
Video2	192	14769	1862	143230
Video3	79	6076	2219	170692
Video4	382	29384	3943	303307
Video5	94	7230	3638	279846
Video6	364	28000	3285	252692

The EEG data of each subject were segmented at intervals of 10 ms to obtain a matrix with a dimension of $64 * 10$. After the EEG data were intercepted at the corresponding starting time and end time, 75% of the information was randomly selected as the train set and 25% as the test set. The data volume of the train set and the test set was about 3:1. The final classification results are shown in Tab. 4.

3.3.3 LSTM-CNN

Both the CNN and LSTM are extensions of traditional neural networks with independent feature extraction and classification functions. The CNN obtains the complete data from the local information aggregation in space, extracting the hierarchical input information. However, LSTM has a sequence in the time dimension that considers the previous input information, displaying prior memory functionality. When the LSTM and CNN are connected, feature fusion can consider

both leads in the space and the time dimensions. Therefore, LSTM-CNN represents the bimodal feature fusion model.

Table 4: Facial-LSTM model classification results

Experimenter	Accuracy (%)
Subject1	85.01
Subject2	76.95
Subject3	94.32
Subject4	90.93
Subject5	94.39
Subject6	94.35
Subject7	88.47
Subject8	90.93
AVG	89.42

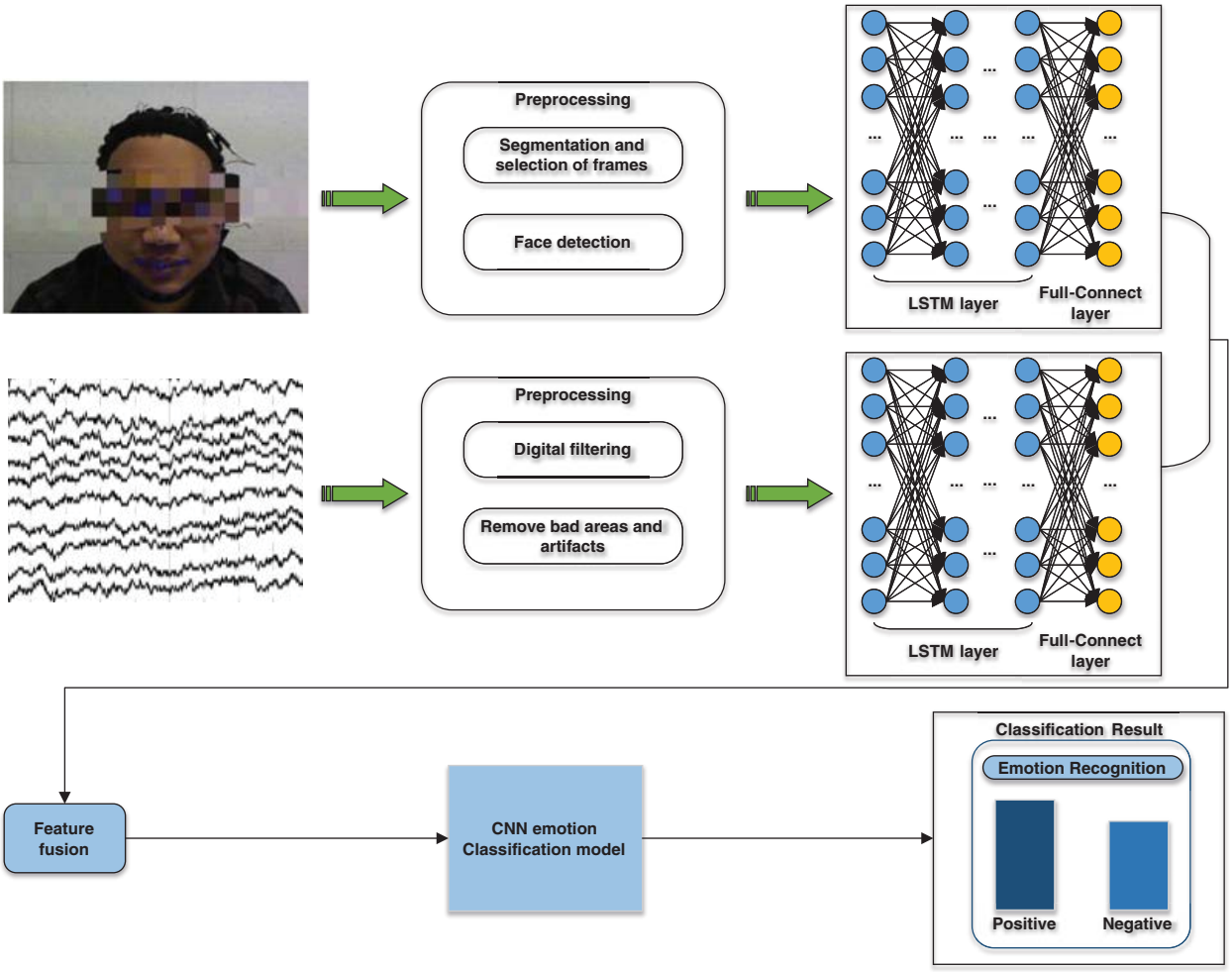

Figure 7: LSTM-CNN model

An output marker appeared in the LSTM-CNN model when the signal entered the LSTM. This tag contained both the original tag information and the related information before the output. Subsequently, the CNN searched for local features via the output-rich feature representation to enhance the accuracy.

During the network structure design of LSTM-CNN, the data input was considered the starting point. The expression and EEG features were extracted using the LSTM model. They were then connected to feature vectors in the output section and sent to the CNN model for classification, as shown in Fig. 7.

This fusion method better utilized the temporal information of each mode, obtaining the characteristic information of its spatial dimension. The classification accuracy of the final LSTM-CNN model was 93.13%.

3.4 Results and Analysis

Tab. 5 compares the classification results of the EEG-LSTM, Facial-LSTM, and LSTM-CNN models:

Table 5: Comparison of the classification results

Model	Average accuracy (%)
EEG-LSTM	86.48
Facial-LSTM	89.42
LSTM-CNN	93.13

This comparison shows that the classification rates of the EEG-LSTM, Facial-LSTM, and LSTM-CNN models are increasing. Therefore, it is feasible to use emotional-modal-assistant EEG signals for emotion recognition. The Facial-LSTM model intercepted EEG signals via the keyframe of facial expression changes, achieving excellent classification. The LSTM-CNN model used the correlation between features for fusion, obtaining the best classification result of the three models.

4 Discussion

This paper aimed to combine expression with EEG data to realize and improve the classification of emotion. Consequently, the Facial-LSTM and LSTM-CNN models were established.

The Facial-LSTM model involved EEG data processing in the first and last frames of the facial expression change output by the model-facial expression recognition system. The LSTM-CNN model fed the preprocessed facial expressions and EEG data into the LSTM for feature extraction, after which the output features were fused and sent to the CNN for classification. The classification accuracy of the Facial-LSTM model was 89.42%, while that of the LSTM-CNN model was 93.13%, improving the precision of the latter model. The results indicated that the bimodal EEG emotion recognition effect surpassed that of single-mode EEG.

5 Conclusion

In recent years, the requirements for human-computer interaction have been increasing. Therefore, accurate identification of human emotions via brain-computer interfaces is essential in providing a bridge during these exchanges [23].

Although current EEG research has become increasingly mature, the moment of emotion generation remains difficult to determine [24]. The expression denotes one of the modes that accurately represent emotion in daily life, being feature-rich features and easy to obtain. Therefore, it is feasible to classify and identify emotions by combining expression and EEG data.

Moreover, there is a correlation between synchronous EEG and facial features, even though they are denoted by different modes. The research regarding bimodal emotion recognition based on EEG and expression indicates that an enhanced effect can be achieved should their feature correlation and the integration of EEG and facial features be better utilized. Therefore, multimodal feature fusion requires further in-depth examination [25].

Acknowledgement: The author thanks all subjects who participated in this research and the technical support from FISTAR Technology Inc.

Funding Statement: This work was supported by the National Nature Science Foundation of China (No. 61503423, H. P. Jiang). The URL is http://www.nsfc.gov.cn/.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. B. Li, J. Sun, Z. B. Xu and L. M. Chen, "Multi-modal 2D + 3D facial expression recognition with deep fusion convolutional neural network," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2816–2831, 2017.

[2] T. Li, L. Y. Wang, Y. Chen, Y. J. Ren, L. Wang *et al.,* "A face recognition algorithm based on LBP-EHMM," *Journal on Artificial Intelligence*, vol. 1, no. 2, pp. 61–68, 2019.

[3] Y. D. Chen, L. M. Wang, S. O. Physics and N. N. University, "Convolutional neural network for face recognition," *Journal of Northeast Normal University (Natural Science Edition)*, vol. 48, no. 2, pp. 70–76, 2016.

[4] S. Y. Lu and F. Evans, "Haar wavelet transform based facial emotion recognition," in *Proc. of the 2017 7th Int. Conf. on Education, Management, Computer and Society*, Madrid, Spain, pp. 342–346, 2017.

[5] S. Qin, Z. Z. Zhu, Y. H. Zou and X. W. Wang, "Facial expression recognition based on Gabor wavelet transform and 2-channel CNN," *International Journal of Wavelets, Multiresolution & Information Processing*, vol. 18, no. 2, pp. 2050003, 2020.

[6] S. Rajan, P. Chenniappan, S. Devaraj and N. Madian, "Novel deep learning model for facial expression recognition based on maximum boosted CNN and LSTM," *IET Image Processing*, vol. 14, no. 7, pp. 1373–1381, 2020.

[7] E. Maiorana, "Deep learning for EEG-based biometric recognition," *Neurocomputing*, vol. 410, no. 1, pp. 374–386, 2020.

[8] B. Y. Zhang, H. P. Jiang and L. S. Dong, "Classification of EEG signal by WT-CNN model in emotion recognition system," in *2017 IEEE 16th Int. Conf. on Cognitive Informatics & Cognitive Computing*, Honolulu, Hawaii, USA, pp. 109–114, 2017.

[9] J. X. Zhang and B. O. Hua, "Research on EEG emotion recognition based on CNN," *Modern Computer*, no. 23, pp. 12–16, 2018. https://doi.org/CNKI:SUN:XDJS.0.2018-23-004.

[10] S. Alhagry, A. A. Fahmy and R. A. El-Khoribi, "Emotion recognition based on EEG using LSTM recurrent neural network," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, pp. 4, 2017.

[11] F. J. Ren, M. L. Yu, M. Hu and Y. Q. Li, "Dual-modality video emotion recognition based on facial expression and BVP physiological signal," *Journal of Image and Graphics*, vol. 23, no. 5, pp. 688–697, 2018.

[12] D. E. Shen, S. Lucia, Y. M. Wan, R. Findeisen and R. D. Braatz, "Bimodal emotion recognition based on facial expression and speech," *Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition)*, vol. 38, no. 1, pp. 60–65, 2018.

[13] Y. Shu and Y. P. Guan, "Audio-visual perception-based multi-modal HCI," *Journal of Engineering-Joe*, vol. 2018, no. 4, pp. 190–198, 2018.

[14] Y. R. Huang, "Fusion of facial expressions and EEG for multi-modal emotion recognition," *Computational Intelligence and Neuroscience*, vol. 2017, pp. 16, 2017.

[15] C. Zhu, Y. K. Wang, D. B. Pu, M. Qi, H. Sun *et al.,* "Multi-modality video representation for action recognition," *Journal on Big Data*, vol. 2, no. 3, pp. 95–104, 2020.

[16] H. Liu and X. Zhou, "Multi-focus image region fusion and registration algorithm with multi-scale wavelet," *Intelligent Automation & Soft Computing*, vol. 26, no. 6, pp. 1493–1501, 2020.

[17] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Networks," *Official Journal of the International Neural Network Society*, vol. 18, no. 5, 6, pp. 602–610, 2005.

[18] H. Jiang, Z. Wang, R. Jiao and S. Jiang, "Picture-induced EEG signal classification based on CVC emotion recognition system," *Computers, Materials & Continua*, vol. 65, no. 2, pp. 1453–1465, 2020.

[19] L. S. Yao, J. W. Zhang, B. Fang, S. L. Zhang, H. Zhou *et al.,* "Design and implementation of facial expression recognition based on LBP and SVM," *Journal of Guizhou Normal University (Natural Sciences)*, vol. 38, no. 1, pp. 63–72, 2020.

[20] M. A. Asghar, M. J. Khan, Fawad, Y. Amin, M. Rizwan *et al.,* "EEG-based multi-modal emotion recognition using bag of deep features: An optimal feature selection approach," *Sensors*, vol. 19, no. 23, pp. 5218, 2019.

[21] M. Imani and G. A. Montazer, "A survey of emotion recognition methods with emphasis on e-learning environments (review)," *Journal of Network and Computer Applications*, vol. 147, pp. 1–40, 2019.

[22] X. M. Cao, Y. H. Zhang, M. Pan, S. Zhu and H. L. Yan, "Research on student engagement recognition method from the perspective of artificial intelligence: Analysis of deep learning experiment based on a multi-modal data fusion," *Journal of Distance Education*, vol. 37, no. 1, pp. 32–44, 2019.

[23] M. Jiang, M. Hu, X. Wang, F. Ren and H. Wang, "Dual-modal emotion recognition based on facial expression and body posture in video sequences," *Laser & Optoelectronics Progress*, vol. 55, no. 7, pp. 167–174, 2018.

[24] Y. J. Li, J. J. Huang, H. Y. Wang and N. Zhong, "Study of emotion recognition based on fusion multi-modal bio-signal with SAE and LSTM recurrent neural network," *Journal on Communications*, vol. 38, no. 12, pp. 109–120, 2017.

[25] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi and C. Fookes, "Deep Spatio-temporal features for multi-modal emotion recognition," in *2017 IEEE Winter Conf. on Applications of Computer Vision*, Santa Rosa, California, pp. 1215–1223, 2017.