Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 8795842, 6 pages
https://doi.org/10.1155/2022/8795842

Research Article

Mental State Assessment in College English Teaching Courses Based on Deep Learning

Beibei Ji

School of Humanities and Foreign Languages, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Correspondence should be addressed to Beibei Ji; jibeibei@xupt.edu.cn

Received 16 June 2022; Accepted 20 July 2022; Published 8 August 2022

Academic Editor: Baiyuan Ding

College English teaching aims at students with different foundations and characteristics. Also, it is necessary to grasp the mental state of students in time during the teaching process. Based on the bimodal information of facial expressions and speech of classroom students, this paper designs a mental state assessment method based on deep learning. For bimodal emotion recognition of facial expressions and speech, a feature fusion method based on sparse canonical correlation analysis (SCCA) is proposed in this paper. First, the emotional features of the facial expression and speech are extracted, respectively. Then, SCCA is used to fuse the emotional features of the two modalities. Finally, the sparse representation-based classification (SRC) is used as the classifier for emotional prediction. Based on the prediction results, the mental state of different students can be grasped, so as to adjust the teaching strategy in a targeted manner. Experiments are carried out based on public datasets. First, the proposed method achieves the average classification accuracy of 92.4%, which is higher than those from the present methods for comparison. Second, under the condition of noise corruption, the proposed method keeps the superior robustness over the comparison methods. The results show that the proposed bimodal emotion recognition method based on SCCA and SRC can achieve higher recognition rates than some present methods.

1. Introduction

With the development of the economy and society, people begin to pay more attention to their own mental health. In the university stage, due to the rapid changes in the internal and external environment of students, many of them cannot adapt in time, which is prone to psychological problems. According to research statistics, the psychological problems of college students have obvious stage characteristics, and many students cannot detect their own psychological changes in time, which leads to the deterioration of psychological problems and has serious consequences. Taking English teaching in colleges and universities as an example, due to the different foundations of English learning and different interests in the student group, different students may have different psychological emotions towards this course. Therefore, it is an important way to correctly adjust the educational method and improve the teaching efficiency to do a good job of mental state assessment in the classroom in time. It is in this context that this paper aims to support the English teaching based on the emotional analysis of students' expressions and speeches during the English teaching classes to obtain their corresponding mental states [1–3].

In the past few decades, many researchers all over the world have used several commonly used human-computer interaction modalities (such as facial expressions, speech, and gestures) to conduct emotion recognition. Some effective expression recognition, speech emotion recognition, and action gesture recognition methods were developed [4–7]. Among these methods, the vast majority are emotion recognition methods based on a single modality. However, in future applications such as emotional robots, unmanned driving, and intelligent transportation, the emotional interaction between humans and machines is often based on bimodal or multimodality. For example, emotional robots can perform human-computer interaction through both facial expression and speech [8–10]. Therefore, the study of

multimodal emotion recognition methods based on facial expressions and speech plays a very important role in the development and progress of future technologies such as emotional robots, unmanned driving, and intelligent transportation [11–13].

At present, some researchers have initially carried out research on bimodal emotion recognition based on facial expressions and speech. Reference [13] proposed a multimodal method of facial expression and speech based on the fusion mechanism of the decision-making layer and achieved certain fusion effects. Reference [14] proposed a multimodal facial expression and speech method based on the multistream hidden Markov model (MHMM), which achieved good performance after fusion. Reference [15] studied a multimodal approach to facial expression and speech based on direct fusion and back-propagation (BP) neural network classifiers. Reference [16] employed the kernel cross-modal factor analysis (KCFA) algorithm to perform feature dimensionality reduction and feature fusion for speech and facial expression modalities. Reference [17] used the kernel entropy component analysis (KECA) algorithm and the decision layer fusion to study facial expression and speech bimodal emotion recognition and achieved high dual-mode emotion recognition on two commonly used emotion databases. In recent years, deep learning models have provided powerful tools for face and speech emotion recognition, and representative methods include convolution neural network (CNN), long short-term memory (LSTM), and generative adversarial network (GAN). [18, 19].

In order to effectively improve the emotion recognition performance during English teaching classes so as to grasp the mental state of different students, this paper proposed a dual-modal emotion analysis method based on facial expression and speech. First, the proposed method extracts the features of facial expression and speech, respectively. Then, the sparse canonical correlation analysis (SCCA) [20–22] algorithm is used to fuse the two kinds of features to obtain a unified feature. Finally, the sparse representation-based classification (SRC) is used for bimodal emotion recognition. Based on the emotion analysis results, each student in the classroom can be observed, and their mental states toward this class can be analyzed. The main contribution of this paper can be summarized as follows. The dual-modal information during the teaching is properly analyzed and fused by SCCA. The fused result can better convey the mental state of the students. SRC performs as the classification scheme and finally gets the result of the mental state. Experiments are performed based on the public dataset. According to the experimental results, the validity and superiority of the proposed method can be verified by comparison with several published methods in this field.

2. Expression and Speech Emotion Feature Extraction

For facial expression and speech bimodal emotion data, this paper first performs feature extraction. For facial expression modalities, this paper uses scale-invariant feature transform (SIFT) to extract features. SIFT feature is a very effective image feature extraction method, which is widely used in action recognition, motion detection, and facial expression recognition, due to its robustness and antinoise advantages.

The previous works show that the SIFT feature extraction process generally includes the extraction of extreme points in the original image, the selection of feature points, the gradient solution of feature points, and the generation of feature point descriptors. In practical applications such as facial expression recognition, the following steps of solving the gradient of feature points and generating step of feature point descriptors are the most critical. The gradient of the feature point is mainly calculated by the following two equations:

$$\rho(x, y) = \sqrt{[\widehat{J}(x, y+1) - \widehat{J}(x, y-1)]^2 + [\widehat{J}(x+1, y) - \widehat{J}(x-1, y)]^2}, \tag{1}$$

$$\theta(x, y) = \tan^{-1}\left\{\frac{\widehat{J}(x, y+1) - \widehat{J}(x, y-1)}{\widehat{J}(x+1, y) - \widehat{J}(x-1, y)}\right\}. \tag{2}$$

In equations (1) and (2), $\rho(x, y)$, $\theta(x, y)$, and $\widehat{J}(x, y)$ represent the required gradient size, gradient direction, and Gaussian smoothed image, respectively. After the gradient is obtained via equations (1) and (2), the direction of each feature point is first calculated by the histogram method, and then, SIFT feature vectors of different dimensions are extracted based on the obtained direction.

For speech modalities, this paper uses openSMILE software to extract the emotional features of speech modalities. Compared with the traditional speech emotion feature extraction method, the speech emotion feature based on openSMILE software is more convenient and direct. As long as the corresponding speech audio is input, the rich speech emotion feature can be directly extracted through a simple operation.

3. Classroom Psychological Assessment Based on Emotion Analysis

3.1. Feature Fusion. At first, SCCA is used to fuse the features of the two modalities from facial expression and speech [20–22]. The SCCA algorithm can be expressed as follows:

$$\underset{A_W,A_S}{\operatorname{argmin}} L_{\mathrm{SCCA}} = \underset{A_W,A_S}{\operatorname{argmin}} \left\| \left(WW^{\mathrm{T}}\right)^{-1/2} \left(W - A_W A_S^{\mathrm{T}} S\right) \right\|_F^2 + \eta_W \|A_W\|_1 + \eta_S \|A_S\|_1. \tag{3}$$

In equation (3), S is the extracted speech feature matrix; W is the facial expression feature matrix extracted by the SIFT method; A_W is the projection matrix of W; A_S is the projection matrix of S; η_W is the sparse parameter of A_W; and η_S is the sparse parameter of A_S.

According to the relevant researches, the problem in equation (3) can be resolved by the augmented Lagrangian algorithm. Let $\widetilde{A}_W = A_W$ and $\widetilde{A}_S = A_S$, equation (3) can be reformulated as follows:

$$\begin{aligned} \underset{A_W,A_S,\widetilde{A}_W,\widetilde{A}_S}{\operatorname{argmin}} L_{\mathrm{SCCA}} = \underset{A_W,A_S,\widetilde{A}_W,\widetilde{A}_S}{\operatorname{argmin}} & \left\| \left(WW^{\mathrm{T}}\right)^{-1/2} \left(W - A_W A_S^{\mathrm{T}} S\right) \right\|_F^2 + \eta_S \|A_S\|_1 + \operatorname{tr}\left[\Gamma_W^{\mathrm{T}} \left(\widetilde{A}_W - A_W\right) + \eta_W \|A_W\|_1\right] \\ & + \operatorname{tr}\left[\Gamma_S^{\mathrm{T}} \left(\widetilde{A}_S - A_S\right)\right] + \frac{\lambda_W}{2} \left\|\widetilde{A}_W - A_W\right\|_F^2 + \frac{\lambda_S}{2} \left\|\widetilde{A}_S - A_S\right\|_F^2, \end{aligned} \tag{4}$$

where λ_W is the norm coefficient of $\widetilde{A}_W - A_W$; λ_S is the norm coefficient of $\widetilde{A}_S - A_S$; Γ_W^T is the Lagrangian multiplier matrix of $\widetilde{A}_W - A_W$; Γ_S^T is the Lagrangian multiplier matrix of $\widetilde{A}_S - A_S$.

Equation (4) can be further rewritten as follows:

$$\begin{aligned} \underset{A_W,A_S,\widetilde{A}_W,\widetilde{A}_S}{\operatorname{argmin}} L_{\mathrm{SCCA}} = \underset{A_W,A_S,\widetilde{A}_W,\widetilde{A}_S}{\operatorname{argmin}} & \operatorname{tr}\left\| \left(WW^{\mathrm{T}}\right)^{-1/2} WW^{\mathrm{T}} \left(WW^{\mathrm{T}}\right)^{-1/2} \right\| - 2\operatorname{tr}\left[\left(WW^{\mathrm{T}}\right)^{-1/2} WS^{\mathrm{T}} \widetilde{A}_S \widetilde{A}_W^{\mathrm{T}} \left(WW^{\mathrm{T}}\right)^{-1/2} \right] \\ & + \operatorname{tr}\left[\left(WW^{\mathrm{T}}\right)^{-1/2} \widetilde{A}_W \widetilde{A}_S^{\mathrm{T}} SS^{\mathrm{T}} \widetilde{A}_S \widetilde{A}_W^{\mathrm{T}} \left(WW^{\mathrm{T}}\right)^{-1/2} \right] + \eta_S \|A_S\|_1 + \eta_W \|A_W\|_1 + \operatorname{tr}\left(\Gamma_S^{\mathrm{T}} \widetilde{A}_S - \Gamma_S^{\mathrm{T}} A_S\right) \\ & + \frac{\lambda_S}{2} \operatorname{tr}\left(\widetilde{A}_S \widetilde{A}_S^{\mathrm{T}} - 2\widetilde{A}_S A_S^{\mathrm{T}} + A_S A_S^{\mathrm{T}} \right) + \frac{\lambda_W}{2} \operatorname{tr}\left(\widetilde{A}_W \widetilde{A}_W^{\mathrm{T}} - 2\widetilde{A}_W A_W^{\mathrm{T}} + A_W A_W^{\mathrm{T}} \right) + \operatorname{tr}\left(\Gamma_W^{\mathrm{T}} \widetilde{A}_W - \Gamma_W^{\mathrm{T}} A_W\right). \end{aligned} \tag{5}$$

Taking the partial derivatives of (5) with respect to $\widetilde{A}_W$ and $\widetilde{A}_S$, respectively, and setting them to 0, we can get

$$\begin{aligned} \frac{\partial L_{\mathrm{SCCA}}}{\partial \widetilde{A}_W} &= -2\left(WW^T\right)^{-1} WS^T \widetilde{A}_S + \Gamma_W + \lambda_W \widetilde{A}_W - \lambda_W A_W + 2\left(WW^T\right)^{-1} \widetilde{A}_W \widetilde{A}_S^T SS^T \widetilde{A}_S = 0, \\ \frac{\partial L_{\mathrm{SCCA}}}{\partial \widetilde{A}_S} &= -2\left(WW^T\right)^{-1} SW^T \widetilde{A}_W + \Gamma_S + \lambda_S \widetilde{A}_S - \lambda_S A_S + 2\left(WW^T\right)^{-1} SS^T \widetilde{A}_S \widetilde{A}_W^T \widetilde{A}_W = 0. \end{aligned} \tag{6}$$

Then, the solution is

$$\begin{aligned} \widetilde{A}_W &= A_W + \frac{2}{\lambda_W}\left(WW^T\right)^{-1} WS^T \widetilde{A}_S - \frac{2}{\lambda_W}\left(WW^T\right) \widetilde{A}_W \widetilde{A}_S^T SS^T \widetilde{A}_S - \frac{\Gamma_W}{\lambda_W}, \\ \widetilde{A}_S &= A_S + \frac{2}{\lambda_S}\left(WW^{\mathrm{T}}\right)^{-1} SW^{\mathrm{T}} \widetilde{A}_W - \frac{2}{\lambda_S}\left(WW^{\mathrm{T}}\right) SS^{\mathrm{T}} \widetilde{A}_S \widetilde{A}_W^{\mathrm{T}} \widetilde{A}_W - \frac{\Gamma_S}{\lambda_S}. \end{aligned} \tag{7}$$

According to the method introduced in [21], after obtaining $\widetilde{A}_W$ and $\widetilde{A}_S$, the solution of A_W and A_S can be transformed into

$$\begin{aligned}&\underset{A_W}{\operatorname{argmin}}\ \frac{\eta_W}{\lambda_W}\|A_W\| + \frac{1}{2}\left\|A_W - \left(\widetilde{A}_W + \frac{\Gamma_W}{\lambda_W}\right)\right\|_F^2,\\&\underset{A_S}{\operatorname{argmin}}\ \frac{\eta_S}{\lambda_S}\|A_S\| + \frac{1}{2}\left\|A_S - \left(\widetilde{A}_S + \frac{\Gamma_S}{\lambda_S}\right)\right\|_F^2.\end{aligned} \tag{8}$$

Then, we can get

$$\begin{aligned}A_W &= \xi_{\frac{\eta_W}{\lambda_W}}\left[\widetilde{A}_W + \frac{\Gamma_W}{\lambda_W}\right],\\A_S &= \xi_{\frac{\eta_S}{\lambda_S}}\left[\widetilde{A}_S + \frac{\Gamma_S}{\lambda_S}\right].\end{aligned} \tag{9}$$

Among them, the function ξ is the threshold function defined in [21].

Finally, the bimodal feature fusion based on SCCA is obtained as follows:

$$\begin{pmatrix} A_W^{\mathrm{T}} W \\ A_S^{\mathrm{T}} S \end{pmatrix}. \tag{10}$$

3.2. Emotion Classification. SRC uses sparse representation for pattern recognition problems, which characterizes the unknown input through training samples of known categories and then determines the category of test samples according to the reconstruction errors of different categories. Supposing $D = [D^1, D^2, \ldots, D^C] \in \mathrm{R}^{d\times N}$ is a global dictionary, where $D^i \in R^{d\times N_i}$ $(i = 1, 2, \ldots, C)$ represents N_i training samples from ith class, and the sparse representation process of the test sample y can be built as follows:

$$\begin{aligned}&\widehat{x} = \underset{x}{\operatorname{argmin}}\ \|x\|_0,\\&\text{s.t.}\|y - Dx\|_2^2 \le \varepsilon.\end{aligned} \tag{11}$$

In the above equations, x represents the sparse coefficient vector. According to relevant works, the algorithms commonly used to solve sparse representation problems include ℓ_1 norm optimization and orthogonal matching pursuit algorithm (OMP) [13–16]. Based on the solution of $\widehat{x}$, the reconstruction errors for the test samples are calculated according to the categories, and finally, the category of the test samples is determined as follows:

$$\begin{aligned}r(i) &= \left\|y - D_i x_i\right\|_2^2 (i = 1, 2, \ldots, C),\\\text{identity}(y) &= \arg\min_i (r(i)),\end{aligned} \tag{12}$$

where x_i is the coefficient vector corresponding to the ith class; $r(i)$ is the corresponding reconstruction error.

Compared with CNN, the classification mechanism of SRC is relatively less dependent on the number of test samples. At the same time, existing research results show that SRC has certain adaptability to complex situations such as noise interference and occlusion. Figure 1 shows the basic process of the proposed method in this paper. Based on the extracted facial expressions and speech features, SCCA is used for feature fusion, and finally, the category is confirmed based on SRC to obtain the emotional and mental state of the current test sample. Specifically, during English teaching courses, the proposed method can be used to automatically obtain the current emotional state of different students, reflecting their psychological acceptance of the current course. Such analysis results can be used to assist in the adjustment and optimization of teaching methods and teaching forms.

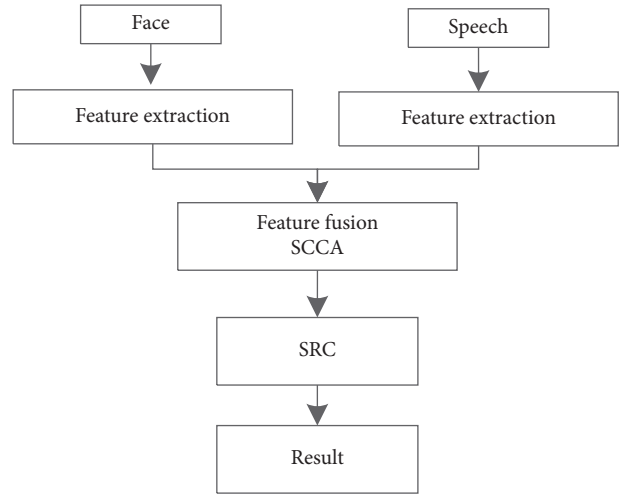

Figure 1: Basic flowchart of the proposed method.

4. Experiment and Analysis

4.1. Introduction of Dataset. The RECOLA dataset is a commonly used dataset for emotion recognition containing speech and visual data, providing audio and video recordings, image, and sound features, some time-specific events, and some other metadata of 46 different experimental participants. The voice module in the dataset contains the original recording, the start and end times of speech, the predicted probability of voice activity, and other characteristics of the voice. The image module in the dataset contains the original videos, the corresponding time of each frame in the video, the predicted probability of face detection, and the features of the image. The original footage was captured by a Logitech webcam, 1080×720 pixels, YUV color mode, and fixed FPS at 25 frames per second. The raw data are annotated with emotion (Arousal and Valence) and type of laughter (silent laughter, normal laughter, talking, and talking laughter). The dataset also provides some other information, such as physiological signals, age, gender, and native language.

In the specific use of this paper, we selected 1200 adult facial expression and speech data samples and selected a fivefold cross-validation strategy for testing. In the process, in order to reflect the performance of the proposed method, several categories of methods from the existing literature are used for comparison, including SVM, CNN, and LSTM. At the same time, the performance test is also carried out under

Table 1: Comparison of performance of different methods.

Method	Average classification accuracy (%)
Proposed	92.4
SVM	85.7
CNN	89.3
LSTM	89.5
Face only	87.1
Speech only	86.2

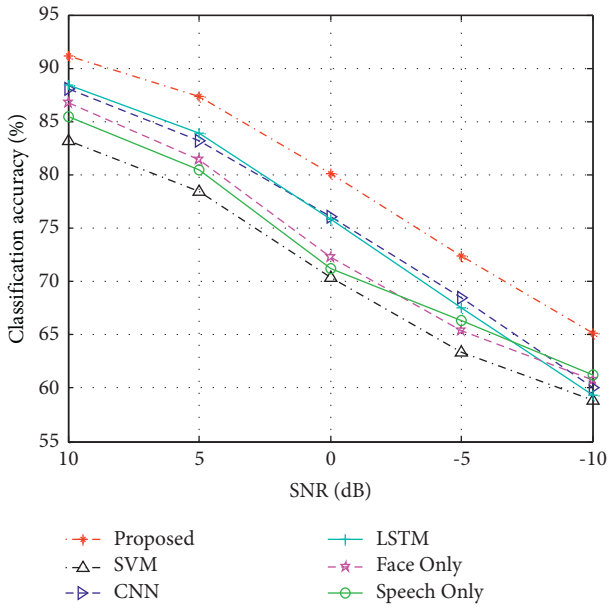

Figure 2: Classification accuracy of different methods under noises.

the condition of facial expression and voice single mode, which is also used as a comparative experiment for the method in this paper. The average classification accuracy is used as a quantitative evaluation index for the performance of various methods, which is the proportion of correctly classified samples to all test samples.

5. Result and Discussion

Based on the constructed dataset, the proposed method and the comparison methods are tested. The statistical results are shown in Table 1. It can be seen that the method in this paper has a performance advantage over the five comparison methods, and the average classification accuracy is at the highest level. Compared with the traditional machine learning methods such as SVM, deep learning models such as CNN and LSTM show more advantageous performance. Compared with the results under the single-modal condition of facial expression and speech, this paper significantly improves the classification accuracy by combining the two modalities, showing the effectiveness of the proposed method.

In the actual process, both the facial expression image and the voice signal may be disturbed by noise. Therefore, it is necessary to improve the ability of emotion analysis under the condition of noise interference. To this end, this paper uses the form of signal-to-noise ratio (SNR) to measure the noise level in the two modalities of facial expression and speech and verifies the classification ability of various methods under the same noise conditions. Figure 2 shows the classification accuracy of different methods with the changing of SNR. It can be seen that the method in this paper maintains the optimal classification performance under various noise levels, showing its robustness.

6. Conclusion

Aiming at the problem of psychological state monitoring in the process of English teaching in colleges and universities, this paper designs an intelligent evaluation method based on emotion recognition. First, the emotional features of facial expression and speech modalities are extracted from the facial expression and speech bimodal emotional database, respectively. Then, the SCCA algorithm is used to fuse the emotional features of the two modalities, and finally, SRC is used to perform emotion recognition. The experimental results show that the dual-modal emotion recognition method based on SCCA and SRC proposed in this paper can achieve a higher recognition rate than some present methods in the same field. The proposed method can carry out targeted analysis for different students and then adjust teaching strategies according to the statistical results, so as to improve the overall quality of English teaching. In the future, the research will be deepened in two ways. First, more available information will be used besides the facial expression and voices. They can be combined to achieve more reliable results. Second, some new intelligent algorithms can be used to process these modalities to further improve the effectiveness and efficiency.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work received Shaanxi Social Science Fund Project: Research on prose creation of contemporary Women writers (2019J016) (in research).

References

[1] H. H. Tu, "A study on the construction of emotion recognition based on multimodal information fusion in English learning cooperative and competitive mode," *Frontiers in Psychology*, vol. 12, Article ID 767844, 2021.

[2] N. Hu, S. Li, L. Li, and H. Xu, "The educational function of English children's movies from the perspective of multiculturalism under deep learning and artificial intelligence," *Frontiers in Psychology*, vol. 12, Article ID 759094, 2021.

[3] M. Du and Y. Qian, "Application of massive open online course to grammar teaching for English majors based on deep

learning," *Frontiers in Psychology*, vol. 12, Article ID 755043, 2021.

[4] J. Yan, X Wang, W Gu, and L. Ma, "Speech emotion recognition based on sparse representation," *Archives of Acoustics*, vol. 38, no. 4, pp. 465–470, 2013.

[5] H. Gunes and M. Piccardi, *From Mono-Modal to Multi-Modal: Affect Recognition Using Visual Modalities*, pp. 154–182, Springer, London UK, 2009.

[6] Y. Jingjie, Z. Wenming, X. Qinxu, G. Lu, H. Li, and B. Wang, "Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1319–1329, 2016.

[7] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[8] A. Zadeh, R. Zellers, E. Pincus, and L. P Morency, "Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[10] H. Yang, C. Yuan, B. Li et al., "Asymmetric 3D convolutional neural networks for action recognition," *Pattern Recognition*, vol. 85, pp. 1–12, 2019.

[11] S. Zhang, X. Zhao, and Q. Tian, "Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 680–688, 2019.

[12] Y. Xing, G. Di Caterina, and J. Soraghan, "A new spiking convolutional recurrent neural network (SCRNN) with applications to event-based hand gesture recognition," *Frontiers in Neuroscience*, vol. 14, Article ID 590164, 2020.

[13] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," *Procedia Computer Science*, vol. 175, pp. 689–694, 2020.

[14] X. Zhao and S. Zhang, "A review on facial expression recognition: feature extraction and classification," *IETE Technical Review*, vol. 33, no. 5, pp. 505–517, 2016.

[15] J. Chen, X. Liu, P. Tu, and A Aragones, "Learning person-specific models for facial expression and action unit recognition," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1964–1970, 2013.

[16] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 3, pp. 529–545, 2016.

[17] G. Yolcu, I. Oztel, S. Kazan et al., "Facial expression recognition for monitoring neurological disorders based on convolutional neural network," *Multimedia Tools and Applications*, vol. 78, no. 22, Article ID 31581, 2019.

[18] N. Sun, Q. Li, R. Huan, J. Liu, and G Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognition Letters*, vol. 119, pp. 49–61, 2019.

[19] D. Liang, H. Liang, Z. Yu, and Y Zhang, "Deep convolutional BiLSTM fusion network for facial expression recognition," *The Visual Computer*, vol. 36, no. 3, pp. 499–508, 2020.

[20] J. Liu, S. Chen, L. Wang et al., "Multimodal Emotion Recognition with Capsule Graph Convolutional Based Representation Fusion," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6339–6343, IEEE, Toronto Canada, June2021.

[21] M. Ren, X. Huang, X. Shi, and W Nie, "Interactive multimodal attention network for emotion recognition in conversation," *IEEE Signal Processing Letters*, vol. 28, pp. 1046–1050, 2021.

[22] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Computation*, vol. 32, no. 5, pp. 829–864, 2020.