Multi-Modal Emotion Recognition by Fusing Correlation Features of Speech-Visual

Chen Guanghui and Zeng Xiaoping

***Abstract*—To effectively fuse speech and visual features, this letter proposes a multi-modal emotion recognition method by fusing correlation features of speech-visual. Firstly, speech and visual features are extracted by two-dimensional convolutional neural network (2D-CNN) and three-dimensional convolutional neural network (3D-CNN), respectively. Secondly, the speech and visual features is processed by feature correlation analysis algorithm in multi-modal fusion. In addition, the class information of speech and visual features are also applied to the feature correlation analysis algorithm, which can effectively fuse speech and visual features and improve the performance of multi-modal emotion recognition. Finally, support vector machines (SVM) completes the classification of multi-modal speech and visual emotion recognition. Experimental results on the RML, eNTERFACE05, BAUM-1 s datasets show that the recognition rate of our method is higher than other state-of-the-art methods.**

***Index Terms*—Correlation, emotion recongnition, multi-modal, speech-visual.**

I. Introduction

With the rapid development of artificial intelligence, emotion recognition has been applied in many fields [1]–[3], including smart homes [4], travel recommendation systems [5], health monitoring [6], etc. People usually express their emotions through external levels (visual [7], speech [8], gestures, etc.) and internal levels (heart rate, breathing, blood pressure, body temperature, EEG signals [9], [10], etc.), where speech and visual features are widely used in emotion recognition because it is simple and intuitive to construct speech and visual datasets. Since the recognition rate of single-modal speech or visual emotion recognition is too low [11], [12], the research focus of speech-visual emotion recognition has shifted from single-mode to multi-mode [13], [14]. At present, multi-modal emotion recognition method generally adopts different feature extraction methods to extract different feature vectors, and then fuses it in series or in parallel [15]–[17]. In addition, Shah. *et al.* [18] first point out that speech and facial areas have a certain correlation. However, current methods do not fully consider this correlation in multi-modal speech-visual emotion recognition, resulting in a low recognition rate. For example, Nie. *et al.* [19] propose a correlation-based graph convolutional network for emotion recognition, which consider the correlation of the intra-class and inter-class videos, but it does not consider the correlation between speech and visual in the video. Kapoor. *et al.* [20] propose a dual-modal framework based on discriminant correlation analysis for emotion recognition, which consider the correlation between speech and visual in the video. However, the class information is not considered in the correlation analysis. Thus, this letter proposes a multi-modal emotion recognition method by fusing correlation features of speech-visual, which is introduced in detail in *II. Proposed method*. The contributions of this letter are as follows: 1) A multi-modal emotion recognition method based on speech-visual correlation features is proposed to improve the performance of emotion recognition; 2) A feature correlation analysis algorithm containing class information is proposed to effectively fuse speech and visual features; 3) A weighting matrix $\mathbf{K}$ and a new inter-class divergence matrix $\mathbf{S}_b$ are constructed to effectively distinguish similar classes.

Manuscript received January 15, 2021; accepted January 21, 2021. Date of publication January 29, 2021; date of current version March 19, 2021. This work was supported in part by the Chongqing Key Project of Technology Innovation and Application Development under Grant cstc2019jscx-mbdxX0050 and in part by the Graduate Research and Innovation Foundation of Chongqing, China, under Grant CYS20071. The associate editor coordinating the review of this manuscript and approving it for publicationwas Prof. Le Lu. *(Corresponding author: Zeng Xiaoping.)*

The authors are with the College of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China (e-mail: C17358445296@163.com; 1415523904@qq.com).

Digital Object Identifier 10.1109/LSP.2021.3055755

II. Proposed Method

The feature fusion method is very important for multi-modal speech-visual emotion recognition. However, the traditional feature methods do not fully consider the correlation between speech and visual modalities, resulting in low recognition rate. Thus, this letter proposes a multi-modal emotion recognition method by fusing correlation features of speech and visual, which is shown in Fig. 1. It mainly includes three parts: pre-processing, feature learning and multi-modal fusion. Moreover, the class information of speech and visual features are also applied to the feature correlation analysis of multi-modal fusion, which can effectively fuse speech and visual features and improve the performance of multi-modal emotion recognition.

A. Pre-Processing

As can be seen from Fig. 1, the pre-processing mainly completes the extraction of both speech Mel-frequency cepstral coefficients (MFCC) features and visual facial expression area. Firstly, the input video is divided into speech and visual data. Secondly, the visual data is extracted the facial expression area by the CenterFace method [21], because it is an effective face detection method by using the deep learning, which can effectively extract the face expression area from the image. Finally, for speech data, the MFCC features with a size of $64 \times 64 \times 3$ are extracted. Specifically, 64 Mel-filter banks from 20 to 8000 Hz are used, frame length is 25 ms, and frame shift is 10 ms. It is

1070-9908 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Mapua University. Downloaded on August 26,2024 at 12:39:41 UTC from IEEE Xplore. Restrictions apply.

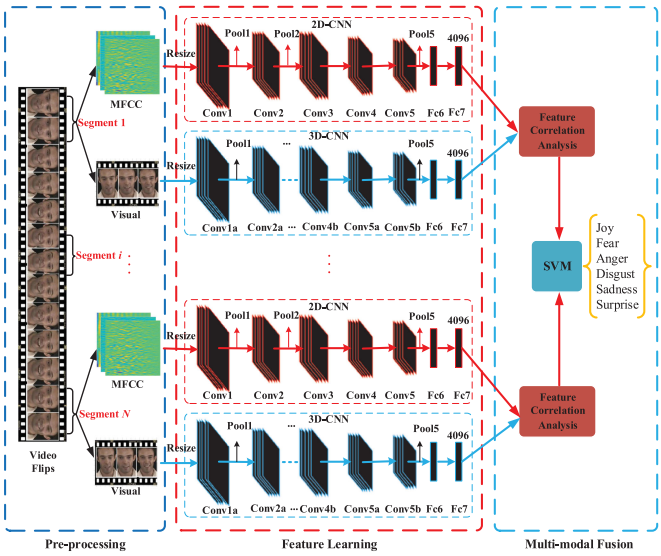

Fig. 1. The framework of our proposed method.

a 30-frame overlap between adjacent segments, and Hamming window is used to smooth. Then a speech MFCC features segment with a size of 64×64 is obtained. The first-order and second-order difference of MFCC are calculated by equation (1).

$$d_t = \frac{\sum_{n=1}^{N} n\left(c_{t+n} - c_{t-n}\right)}{2\sum_{n=1}^{N} n^2} \tag{1}$$

where d_t represents the *t-th* first-order difference. c_{t+n} represents the *(t + n)-th* cepstrum coefficient. N represents the regression window with a typical value of 2.

B. Feature Learning

As can be seen from Fig. 1, the feature learning includes speech feature learning and visual feature learning. The speech feature learning uses the two-dimensional convolutional neural network (2D-CNN), and the visual feature learning uses the three-dimensional convolutional neural network (3D-CNN).

1) Speech Feature Learning: AlexNet [22] is used to initialize 2D-CNN for speech feature learning. It has five convolutional layers (Conv1-Conv2-Conv3-Conv4-Conv5), three maximum pooling layers (Pool1-Pool2-Pool5) and three fully connected (Fc6-Fc7-Fc8) layers. The softmax layer (Fc8) of AlexNet is replaced with the number of emotion categories in this letter. After that, the new softmax layer is used to fine-tune the model, and the back propagation algorithm is used to update the parameters. The training process of 2D-CNN is as follows:

Firstly, the equation (2) is optimized to update speech network by using back propagation algorithm.

$$\min_{\theta^s} \sum_{i=1}^{K} L\left(\text{softmax}\left(\Upsilon^s\left(S_i;\theta^s\right)\right), y_i\right) \tag{2}$$

where θ^S is the parameters of 2D-CNN, S_i represents the speech training data, y_i represents the class label of a segment. Please note that the emotion labels of a video clip is used as the class label of a segment. $\Upsilon^s(S_i;\theta^s)$ represents the output of Fc7 (which is shown in Fig. 1) in 2D-CNN with network parameter θ^s, which has 4096 nodes. The new softmax log-loss is defined as

$$L(S,y) = -\sum_{j=1}^{l} y_j \log\left(y_j^S\right) \tag{3}$$

where y_j represents the *j-th* ground truth label of training data, y_j^S represents the *j-th* output value of the softmax layer, and l represents the total number of class labels.

2) Visual Feature Learning: The C3D-Sports-1M model [23] is used to initialize the 3D-CNN for visual feature learning. It has eight convolutional layers (Conv1a-Conv2a-Conv3a-Conv3b-Conv4a-Conv4b-Conv5a-Conv5a), five maximum pooling layers (Pool1-Pool2-Pool3-Pool4-Pool5), and there fully connected layers (Fc6-Fc7-Fc8). The optimization and training of the visual model are similar to the speech model according to (2) and (3). In addition, to meet the input requirements of 3D-CNN, the size of face images is resized to $150 \times 110 \times 3$.

C. Multi-Modal Fusion

As can be seen from Fig. 1, the multi-modal fusion first uses the correlation analysis algorithm to fuse the speech and visual features. In addition, the correlation analysis also use the class information of speech and visual features, which can effectively fuse speech and visual features and improve the recognition rate of multimodal emotion recognition. After that, the speech and visual features processed by correlation analysis are classified by support vector machines (SVM) with Gaussian kernel function, where the principle of feature correlation analysis algorithm is as follows:

Firstly, the speech and visual sample space are defined as $X = \{x | x \in R^p\}$ and $Y = \{y | y \in R^q\}$, respectively, and there are c pattern classes $\{w_1, w_2, \cdots w_c\}$. The x and y represent the speech and visual feature vectors of the output of Fc7 of 2D-CNN and 3D-CNN in Fig. 1, respectively. Thus, the speech and visual feature training sample set are defined as

$$\begin{cases} \left\{x_{i,j} | x_{i,j} \in X, i = 1, 2, \cdots c, j = 1, 2, \cdots N_{(w_i)}\right\} \\ \left\{y_{i,j} | y_{i,j} \in Y, i = 1, 2, \cdots c, j = 1, 2, \cdots N_{(w_i)}\right\} \end{cases} \tag{4}$$

where $x_{i,j}$ and $y_{i,j}$ represent the *j-th* speech and visual feature training sample of pattern class w_i, respectively. $N_{(w_i)}$ represents the number of samples of pattern class w_i. To avoid the problem of incompatible feature vectors in fusion process, the feature vectors of two sets are standardized by equation (5).

$$\begin{cases} X' = \frac{X - \mu_x}{\sigma_x} \\ Y' = \frac{Y - \mu_y}{\sigma_y} \end{cases} \tag{5}$$

where μ_x and μ_y represent the mean value of feature vector in the sample space, respectively. σ_x and σ_y represent the mean value of standard deviation of feature vector on each component, respectively.

Secondly, the speech and visual feature correlation analysis generally use the canonical correlation analysis (CCA) [24] and [25], because CCA can process two datasets and use correlation as a measure of the similarity of samples in two datasets. However, in the information fusion for the purpose of

Authorized licensed use limited to: Mapua University. Downloaded on August 26,2024 at 12:39:41 UTC from IEEE Xplore. Restrictions apply.

pattern recognition, CCA does not make full use of the class information and ignores the differences between classes. These neglected class information are the most important for pattern recognition. Thus, the class information is added to the feature correlation analysis algorithm to improve the recognition rate of multi-modal emotion recognition. The steps are as follows:

a) A weight matrix $\mathbf{K}$ is constructed to represent the similarity between classes. To make better use of class information, feature extraction should focus on separating class samples with greater similarity to obtain better identification features. Thus, a weight matrix $\mathbf{K}$ is constructed to represent the similarity between classes, where the weight matrix $\mathbf{K}$ is defined as

$$\begin{aligned}\mathbf{K} &= \operatorname{sim}\left(\boldsymbol{v}^{(w_i)}, \boldsymbol{v}^{(w_j)}\right) \\ &= \frac{\sum_{l=1}^{m}\left(\boldsymbol{v}_l^{(w_i)}\boldsymbol{v}_l^{(w_j)}\right)}{\sqrt{\sum_{l=1}^{m}\left(\boldsymbol{v}_l^{(w_i)}\right)^2}\sqrt{\sum_{l=1}^{m}\left(\boldsymbol{v}_l^{(w_j)}\right)^2}}\end{aligned} \tag{6}$$

where $i, j \in [1, c]$, thus, the weight matrix $\mathbf{K}$ is a $c \times c$-dimensional matrix. $\operatorname{sim}(\boldsymbol{v}^{(w_i)}, \boldsymbol{v}^{(w_j)})$ represents the cosine similarity of the average vector of the speech and visual samples of pattern class w_i and w_j. $\boldsymbol{v}^{(w_i)}$ represents the average vector of pattern class w_i, which is defined as

$$\boldsymbol{v}^{(w_i)} = \frac{1}{N^{(w_i)}}\sum_{\rho=1}^{N^{(w_i)}} b_\rho^{(w_i)} \tag{7}$$

where $b_\rho^{(w_i)}$ represents the $\rho - th$ sample vector of pattern class w_i. It can be seen from (6) that the more similar the classes, the larger the corresponding weight.

b) The weight matrix $\mathbf{K}$ is used to optimize the inter-class divergence matrix $\mathbf{S}$ to obtain a new inter-class divergence matrix $\mathbf{S}_b$, where inter-class divergence matrix $\mathbf{S}$ is defined as

$$\mathbf{S} = \sum_{i=1}^{c} N^{(w_i)}\left(\boldsymbol{m}^{(w_i)} - \boldsymbol{m}\right)\left(\boldsymbol{m}^{(w_i)} - \boldsymbol{m}\right)^T \tag{8}$$

where *T* represents the transpose of matrix. $N^{(w_i)}$ represents the number of samples of pattern class w_i. The inter-class divergence matrix $\mathbf{S}$ describes the degree of dispersion between the mean vector $\boldsymbol{m}^{(w_i)}$ of pattern class w_i and the overall mean vector $\boldsymbol{m}$. Moreover, the elements on the diagonal of the matrix $\mathbf{S}$ represent the variance of pattern class w_i relative to the overall average vector $\boldsymbol{m}$. Thus, the inter-class divergence matrix can also be used to describe the degree of dispersion between any two pattern classes. Moreover, to increase the difference between the various classes, the weight matrix $\mathbf{K}$ is used to optimize the inter-class divergence matrix, which is defined as

$$\begin{aligned}\mathbf{S}_b = \sum_{i=1}^{c}\sum_{j=1}^{c} N^{(w_i)}N^{(w_j)}\mathbf{K}\left(\boldsymbol{m}^{(w_i)} - \boldsymbol{m}^{(w_j)}\right) \\ \left(\boldsymbol{m}^{(w_i)} - \boldsymbol{m}^{(w_j)}\right)^T\end{aligned} \tag{9}$$

where $\mathbf{S}_b$ describes the degree of dispersion between classes. The more obvious the difference between classes, the more beneficial it is for multimodal emotion recognition. Thus, the value of $\mathbf{S}_b$ should be large enough. It can be seen from (6) that the more similar the classes, the larger the corresponding weight $\mathbf{K}$. Thus, the weight $\mathbf{K}$ is added to $\mathbf{S}_b$ to effectively distinguish similar classes, which can effectively improve the performance of multimodal emotion recognition.

c) The objective function of CCA add new inter-class divergence constraint $tr[\mathbf{S_b}]$ to construct new objective function, where the objective function of CCA [24] is defined as

$$\begin{cases}\arg\max\limits_{\alpha,\beta}\alpha^T\mathbf{S_{xy}}\beta \\ \alpha^T\mathbf{S_{xx}}\alpha = 1 \\ \beta^T\mathbf{S_{yy}}\beta = 1\end{cases} \tag{10}$$

Thus, to use category information, the objective function of CCA is added with a new inter-class divergence constraint $tr[\mathbf{S_b}]$, which is defined as

$$\begin{cases}\arg\max\limits_{\alpha,\beta}\alpha^T\mathbf{S_{xy}}\beta + \delta tr\left[\mathbf{S_b}\right] \\ \alpha^T\mathbf{S_{xx}}\alpha = 1 \\ \beta^T\mathbf{S_{yy}}\beta = 1\end{cases} \tag{11}$$

The (11) also the new objective function defined by the correlation analysis in this letter, where $\mathbf{S_{xx}}$ and $\mathbf{S_{yy}}$ represent the covariance matrix of training sample space *X* and *Y*. $\mathbf{S_{xy}}$ represents the cross-covariance matrix of training sample space *X* and *Y*. α, β represent a pair of projection vectors. δ is the adjustment factor.

Thirdly, to solve the optimization problem in the (11), a Lagrangian function is defined as

$$\begin{aligned}L = \alpha^T\mathbf{S_{xy}}\beta + \delta tr\left[\mathbf{S_b}\right] - \frac{\lambda_1}{2}\left(\alpha^T\mathbf{S_{xx}}\alpha - 1\right) \\ - \frac{\lambda_2}{2}\left(\beta^T\mathbf{S_{yy}}\beta - 1\right)\end{aligned} \tag{12}$$

The Lagrangian multiplier method in [26] is used to optimize the (12) to obtain the parameter α, β. Based on the parameter α, β, the transformed feature vector is defined as

$$\begin{cases}X'' = \left(\alpha_1^T x, \ldots, \alpha_d^T x\right) = \left(\alpha_1^T, \ldots, \alpha_d^T\right)x = \mathbf{W_x}^T x \\ Y'' = \left(\beta_1^T y, \ldots, \beta_d^T y\right) = \left(\beta_1^T, \ldots, \beta_d^T\right)y = \mathbf{W_y}^T y\end{cases} \tag{13}$$

where α_i, β_i represent the *i-th* pair of projection vectors of *x* and *y*, respectively. $\mathbf{W_x}$ and $\mathbf{W_y}$ represent the projection matrix of *x* and *y*. Thus, the final fusion feature is defined as

$$Z = \begin{pmatrix}\mathbf{W_x}^T x \\ \mathbf{W_y}^T y\end{pmatrix} \tag{14}$$

III. Results and Analysis

RML [27], eNTERFACE05 [28], BAUM-1s [29] datasets were selected to evaluate the performance of our method. The experimental setup is shown in Table I. TenserFlow toolbox is used to implement 3D-CNN and 2D-CNN. LIBSVM toolkit is used to implement SVM by using Gaussian kernel function. In training stage, the 2D-CNN and 3D-CNN are first trained separately, then the speech and visual features output by the Fc7 layer of 2D-CNN and 3D-CNN are processed by the correlation analysis, and finally, the speech and visual data after correlation analysis are used to train SVM.

Authorized licensed use limited to: Mapua University. Downloaded on August 26,2024 at 12:39:41 UTC from IEEE Xplore. Restrictions apply.

TABLE I
EXPERIMENT SETUP

Hardware Platform	CPU	Intel(R)Core(TM)i7-9700K CPU@3.60G
	Memory	32GB
	GPU	NVIDIA GeForce RTX 2080
Training Setup	2D-CNN	3D-CNN
Batch Size	30	30
Number of Epochs	500	500
Dropout Parameter		0.3
Learning Rate		0.001

TABLE II
CLASS INFORMATION COMPARISON EXPERIMENT RESULT

Datasets	Without (%)	With (%)
RML	92.39	96.79
eNTERFACE05	95.12	98.92
BAUM-1s	66.52	71.26

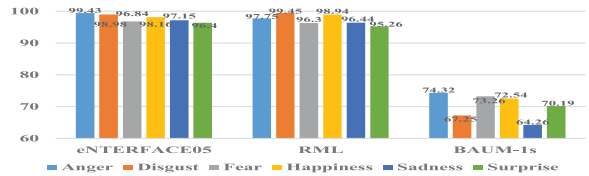

Fig. 2. The recognition rate of each class on eNTERFACE05, RML, and BAUM-1 s datasets, where the unit is %.

A. Class Information Comparison Experiment

The results of the class information comparison experiment are shown in Table II. The recognition rate of multi-modal emotion recognition (96.79% for RML, 98.92% for eNTERFACE05, 71.26% for BAUM-1 s) by using class information is higher than that of not using class information (92.39% for RML, 95.12% for eNTERFACE05, 66.52% for BAUM-1 s). The reason is that the inter-class divergence constraint $tr[\mathbf{S_b}]$ containing class information is added to the correlation analysis of speech and visual features, which can effectively distinguish similar classes, thereby improving the recognition rate of multi-modal emotion recognition.

B. Comparison With the State-of-The-Art

Our method is compared with other state-of-the-art methods on the three datasets, the recognition rate of each class and total are shown in Fig. 2 and Table III, respectively.

It can be seen from Fig. 2 that emotions with higher intensity seem to be easier to recognize. As shown in Table III, the recognition rate of our method is higher than that of the method in [13]–[15] for RML dataset. For eNTERFACE05 dataset, the recognition rate of our method is also higher than that of the method in [13], [15], [19], [20]. For BAUM-1 s dataset, the recognition rate of our method is also higher than that of the method in [13], [14], [17].

The reason is that the method in [13] and [14] did not consider the correlation between speech and visual modalities, only simply connect speech and visual multi-modal features in series or parallel. Thus, the recognition rate of the method in [13] and [14] are much lower than that of other methods.

TABLE III
EXPERIMENT RESULT

Datasets	Refs.	Recognition rate (%)
RML	Zhang. et al. [13]	80.36
	Kansizoglou. et al. [14]	82.97
	Ma. et al. [15]	90.10
	ours	**96.79**
eNTERFACE05	Zhang. et al. [13]	85.97
	Ma. et al. [15]	92.30
	Nie. et al. [19]	97.07
	Kapoor. et al. [20]	98.50
	ours	**98.92**
BAUM-1s	Zhang. et al. [13]	54.57
	Kansizoglou. et al. [14]	56.01
	Ma Fei et al. [17]	67.59
	ours	**71.26**

The method in [15] performs the cross-modal modeling of speech and visual data in the preprocessing. Although it can improve the performance of multimodal emotion recognition by reducing the speech noise and visual redundancy, it does not consider the correlation between speech and visual features. Thus, the recognition rate of the method in [15] is higher than that of the method in [13] and [14], but it is lower than that of other methods. The method in [17] uses speech and visual public information by the correlation analysis to achieve multi-modal emotion recognition. Although it can improve the performance of multimodal emotion recognition by enhancing the stability of features that are learned from different modalities, it did not make full use of the different information of speech and visual features, which is also very important for multimodal emotion recognition. The method in [19] consider the correlation of the intra-class and inter-class videos, but it does not consider the correlation between speech and visual in the video. The method in [20] consider the correlation between speech and visual in the video, but it does not consider the class information in correlation analysis. Thus, the recognition rate of the method in [17], [19] and [20] is also lower than that of our method. Because our method first uses feature correlation analysis algorithm to fully consider the correlation between the speech and visual features. Secondly, the class information is also added the feature correlation analysis algorithm to effectively fuse speech and visual features. In addition, the weighting matrix $\mathbf{K}$ and a new inter-class divergence matrix $\mathbf{S}_b$ are constructed to effectively distinguish similar classes. Thus, our method can effectively improve the recognition rate of multi-modal emotion recognition.

IV. CONCLUSION

Based on the correlation analysis of visual and speech features, this letter proposes a multimodal emotion recognition method. It can effectively distinguish similar classes and fuse speech and visual features, thereby effectively improving the performance of multi-modal emotion recognition. In addition, the recognition rate of multimodal emotion recognition is 96.79% on the RML dataset, 98.92% on the eNTERFACE05 dataset, and 71.26% on the BAUM-1 s dataset, which is higher than other state-of-the-art methods.

Authorized licensed use limited to: Mapua University. Downloaded on August 26,2024 at 12:39:41 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] H. Wei and Z. Zhang, "A survey of facial expression recognition based on deep learning," in *Proc. 15th IEEE Conf. Ind. Electron. Appl.*, Kristiansand, Norway, 2020, pp. 90–94.

[2] N. Hajarolasvadi, M. A. Ramírez, W. Beccaro, and H. Demirel, "Generative adversarial networks in human emotion synthesis: A review," *IEEE Access*, vol. 8, pp. 218499–218529, 2020.

[3] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017.

[4] M. Kissoon Curumsing, N. Fernando, M. Abdelrazek, R. Vasa, K. Mouzakis, and J. Grundy, "Emotion-oriented requirements engineering: A case study in developing a smart home system for the elderly," *J. Syst. Softw.* vol. 147, pp. 215–229, 2019.

[5] K. Lim, Hui, J. Chan, S. Karunasekera, and C. Leckie, "Tour recommendation and trip planning using location-based social media: A survey," *Knowl. Inf. Syst.*, vol. 60, pp. 1247–1275, 2019.

[6] H. Zhao *et al.*, "Adaptive gait detection based on foot-mounted inertial sensors and multi-sensor fusion," *Inf. Fusion* vol. 52, pp. 157–166, 2019.

[7] J. He, C. Zhang, X. He, and R. Dong, "Visual recognition of traffic police gestures with convolutional pose machine and handcrafted features," *Neurocomputing*, vol. 390 pp. 248–259, 2020.

[8] O. Kwon, I. Jang, C. Ahn, and H. Kang, "An effective style token weight control technique for end-to-end emotional speech synthesis," *IEEE Signal Process. Lett.*, vol. 26, no. 9, pp. 1383–1387, Sep. 2019.

[9] N. Cudlenco, N. Popescu, and M. Leordeanu, "Reading into the mind's eye: Boosting automatic visual recognition with EEG signals," *Neurocomputing*, vol. 386, pp. 281–292, 2020.

[10] L. Shen, Z. Liu, and Y. Li, "EEG based dynamic RDS recognition with frequency domain selection and bispectrum feature optimization," *J. Neurosci. Methods*, vol. 337, 2020, Art. no. 108650.

[11] Y. Tian, J. Cheng, Y. Li, and S. Wang, "Secondary information aware facial expression recognition," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1753–1757, Dec. 2019.

[12] W. Zhang, K. Fu, X. Sun, Y. Zhang, H. Sun, and H. Wang, "Joint optimisation convex-negative matrix factorisation for multi-modal image collection summarisation based on images and tags," *IET Comput. Vis.*, vol. 13, no. 2, pp. 125–130, 2018.

[13] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio-visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, Oct. 2018.

[14] I. Kansizoglou, L. Bampis, and A. Gasteratos, "An active learning paradigm for online audio-visual emotion recognition," *IEEE Trans. Affective Comput.*, early access, Dec. 2020, doi: 10.1109/TAFFC.2019.2961089

[15] Y. Ma, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (AVEF): A deep efficient weighted approach," *Inf. Fusion*, vol. 46, pp. 184–192, 2019.

[16] J. Cornejo and H. Pedrini, "Bimodal emotion recognition based on audio and facial parts using deep convolutional neural networks," *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl.*, Boca Raton, FL, USA, 2019, pp. 111–117.

[17] F. Ma, W. Zhang, Y. Li, S.-L. Huang, and L. Zhang, "Learning better representations for audio-visual emotion recognition with common information," *Appl. Sci.*, vol. 10, no. 20, pp. 7239.1–7239.23, 2020.

[18] D. Shah and S. Marshall, "Lip synchronization through alignment of speech and image data," in *Proc. 5th Int. Conf. Image Process. Appl.*, 1995, pp. 598–602.

[19] W. Nie, M. Ren, J. Nie, and S. Zhao, "C-GCN: Correlation based graph convolutional network for audio-video emotion recognition," *IEEE Trans. Multimedia*, early access, Oct. 21, 2020, doi: 10.1109/TMM.2020.3032037.

[20] R. Kapoor, S. Saifi, T. Kapoor, U. Bisaria, and N. Singh, "Dual-modal emotion recognition using discriminant correlation analysis," *Proc. Int. Conf. Electron. Sustain. Commun. Syst.*, Coimbatore, India, 2020, pp. 261–267.

[21] Y. Xu, W. Yan, G. Yang, J. Luo, T. Li, and J. He, "CenterFace: Joint face detection and alignment using face as point," *Sci. Prog.*, vol. 2020, 2020, Art. no. 7845384, doi: 10.1155/2020/7845384.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun.*, vol. 60, no. 6, pp. 84–90, 2017.

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," *Proc. IEEE Int. Conf. Comput. Vis.*, New York, NY, USA, 2015, pp. 4489–4497.

[24] N. Koide-Majima and K. Majima, "Quantum-inspired canonical correlation analysis for exponentially large dimensional data," *Neural Netw.*, vol. 135, pp. 55–67, 2020.

[25] L. Gao, L. Qi, E. Chen, and L. Guan, "Discriminative multiple canonical correlation analysis for information fusion," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1951–1965, Apr. 2018.

[26] K. Shahid Tanzeem and I. D. Schizas, "Unsupervised kernelized correlation-based hyperspectral unmixing with missing pixels," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4509–4520, Jul. 2019.

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

[28] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops.*, 2006, pp. 8–8.

[29] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Trans. Affective Comput.*, vol. 8, no. 3, pp. 300–313, 1 Jul.–Sep. 2017.

Authorized licensed use limited to: Mapua University. Downloaded on August 26,2024 at 12:39:41 UTC from IEEE Xplore. Restrictions apply.