Multimodal Emotion Recognition Using Deep Generalized Canonical Correlation Analysis with an Attention Mechanism

Yu-Ting Lan[1], Wei Liu [1], Bao-Liang Lu[1,2,3,4,*]
[1]Center for Brain-like Computing and Machine Intelligence
Department of Computer Science and Engineering
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
[3]Brain Science and Technology Research Center; [4]Qing Yuan Research Institute
Shanghai Jiao Tong University, Shanghai, 200240, China

***Abstract*—Since multimodal learning is able to take advantage of the complementarity of multimodal signals, the performance of multimodal emotion recognition usually surpasses that based on a single modality. In this paper, we introduce deep generalized canonical correlation analysis with an attention mechanism (DGCCA-AM) to multimodal emotion recognition. This model extends the conventional canonical correlation analysis (CCA) from two modalities to arbitrarily numerous modalities and implements multimodal adaptive fusion with an attention mechanism. By adjusting the weights matrices to maximize the generalized correlation of different modalities, DGCCA-AM extracts emotion-related information from multiple modalities and discards noises. The attention mechanism allows a neural network to learn adaptive fusion weights for different modalities and produces a more effective multimodal fusion and superior emotion recognition performance. We evaluate DGCCA-AM on a public multimodal dataset, SEED-V. Our experimental results demonstrate that DGCCA-AM achieves a state-of-the-art mean accuracy of 82.11% and standard deviation of 2.76% for five emotion classifications with three modalities.**

***Index Terms*—Multimodal deep learning, deep generalized canonical correlation analysis, attention mechanism, multimodal emotion recognition.**

I. Introduction

Emotions are an important part of human life [10]. With the current boom of human-computer interaction (HCI), recent studies have been devoted to enhancing computers with multiple abilities to build better interactions between computers and users. Emotion recognition is essential to HCI because effectively detecting the emotional state makes it possible to build a reliable bridge between computers and users [16].

In recent years, various physiological-signal-based methods have been developed for emotion recognition. These signals are more accurate and difficult to be deliberately changed by users. As a physiological signal that directly reflects brain activity, electroencephalography (EEG) has been demonstrated as a reliable and promising indicator of the human mental state [17]. Kim *et al*. showed that electromyogram, electrocardiogram, skin conductivity, and respiration changes were reliable signals for emotion recognition [19]. Võ *et al*. found that the pupil old/new effect was clearly diminished for emotional words [20].

Since emotions are complex psychophysiological phenomena associated with many nonverbal cues, it is difficult to build robust emotion recognition models using a single modality. Recent research has indicated that multimodal learning is more powerful than unimodal learning [3]. Multimodal learning involves relating information from multiple sources and attempting to utilize the complementary property and independent information of the multimodal signals. Zheng *et al*. combined EEG signals and eye movement signals to build a fusion model and successfully improved the emotion recognition performance [13]. Lu *et al*. employed feature-level concatenation, MAX fusion, SUM fusion, and fuzzy integral fusion to merge EEG and eye movement features and found that EEG and eye movement features have complementary properties for emotion recognition [7].

In recent years, deep neural networks have performed well for classification problems. They are able to learn high-level representation from the raw input features and effectively achieve the classification goal [11]. Qiu *et al*. adopted deep canonical correlation analysis (DCCA) for multimodal emotion recognition and obtained significant performance improvement with respect to three emotion recognition tasks [12]. DCCA computes the representation of multiple modalities by passing them through multiple stacked layers of nonlinear transformation. However, DCCA can only maximize the correlation between two different modalities due to the limitation of the CCA constraint. To extend DCCA from two modalities to arbitrarily numerous modalities, we introduce deep generalized correlation constraints analysis (DGCCA) [4] to multimodal emotion recognition in this paper.

The attention mechanism was initially adopted in the area of image processing to capture the local features of images such as human vision. In recent years, it has been widely used in the fields of natural language processing and computer vision to solve the classification problem. Liu *et al*. combined the attention mechanism with LSTM and successfully alleviated the problem of how to capture precise sentiment expressions

*Corresponding author: Bao-Liang Lu (bllu@sjtu.edu.cn)

in aspect-based sentiment analysis for reasoning [21]. A self attention mechanism that is specific to classification models has been proposed to inform the classifier regarding which parts of the input are more relevant to the output class [9]. In this paper, we introduce the attention mechanism [9] to DGCCA and propose a novel model called DGCCA-AM. Compared with weighted sum fusion in DCCA [3], it assigns each modality an adaptive weight and alleviates the need for users to seek the prior knowledge of the modality.

Our main contributions are as follows:

1) We introduce DGCCA to multimodal emotion recognition and extend DCCA from two modalities to arbitrarily numerous modalities for the first time.
2) We introduce the attention mechanism to DGCCA and propose a novel model called DGCCA-AM.
3) We demonstrate that EEG, eye image (EIG), and eye movement (EYE) modalities have complementary properties.
4) Our new model, DGCCA-AM, achieves considerable improvement in prediction accuracy, 82.11% with a standard deviation of 2.76%, with respect to a three-modality dataset, SEED-V.

The remainder of this paper is organized as follows. Section II summarizes the related work about CCA and DCCA. In Section III, we describe feature extraction methods and the framework of DGCCA-AM. The experimental settings, results, and analysis are presented in Section IV. Finally, the conclusions are given in Section V.

II. Related work

A. Canonical Correlation Analysis

Canonical correlation analysis (CCA) was proposed by Hotelling [1]. It is a standard statistical technique and a fundamental multimodal learning method for finding linear projections of two vectors that are maximally correlated. Hardoon *et al*. introduced CCA to machine learning [18].

Let $X_1 \in \mathbb{R}^{n_1}$ and $X_2 \in \mathbb{R}^{n_2}$ be two vectors, with covariance matrix Σ_{11} for X_1, Σ_{22} for X_2 and cross-variance matrix Σ_{12}. CCA attempts to find the following linear transformations of X_1 and X_2 that maximize the correlation between them.

$$\begin{aligned}(w_1, w_2) &= \underset{w_1 \in \mathbb{R}^{n_1}; w_2 \in \mathbb{R}^{n_2}}{\operatorname{argmax}} \operatorname{corr}\ w_1^\top X_1, w_2^\top X_2 \\ &= \underset{w_1 \in \mathbb{R}^{n_1}; w_2 \in \mathbb{R}^{n_2}}{\operatorname{argmax}} \frac{w_1^\top \Sigma_{12} w_2}{\sqrt{w_1^\top \Sigma_{11} w_1 w_2^\top \Sigma_{22} w_2}}.\end{aligned} \tag{1}$$

Eq. (1) is invariant to linear transformation with two scaling factors w_1 and w_2. We can reformulate the equation as follows:

$$(w_1, w_2) = \underset{w_1^\top \Sigma_{11} w_1 = w_2^\top \Sigma_{22} w_2 = 1}{\arg\max} w_1^\top \Sigma_{12} w_2. \tag{2}$$

To find multiple results of w_1^i, w_2^i, subsequent projections are also constrained to be uncorrelated with previous ones, i.e., $w_1^i \Sigma_{11} w_1^j = w_2^i \Sigma_{22} w_2^j = 0$ for $i < j$. Combining the top k projection vectors w_1^i into a matrix $A_1 \in \mathbb{R}^{n_1 \times k}$ as column vectors and similarly placing w_2^i into $A_2 \in \mathbb{R}^{n_2 \times k}$, we then identify the top $k \leq \min(n_1, n_2)$ projections as follows:

$$\begin{aligned}\text{maximize} \quad & \operatorname{tr}\ A_1^\top \Sigma_{12} A_2 \\ \text{subject to} \quad & A_1^\top \Sigma_{11} A_1 = A_2^\top \Sigma_{22} A_2 = I.\end{aligned} \tag{3}$$

To solve the objective function of Eq. (3), we first define $T = \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2}$, and let U_k and V_k be the matrices of the first k left singular and right singular vectors of T. The optimal objective value is then the sum of the top k singular values of T, and the optimum is obtained at $(A_1, A_2) = \left(\Sigma_{11}^{-1/2} U_k, \Sigma_{22}^{-1/2} V_k\right)$. This method requires the covariance matrices Σ_{11} and Σ_{22} to be nonsingular, which is usually satisfied in practice.

B. Deep Canonical Correlation Analysis for multimodal recognition

Deep canonical correlation analysis (DCCA) was proposed by Andrew and colleagues [2]. DCCA combines the powerful neural network with CCA and overcomes the limitation that CCA can only find a linear transformation of two input vectors. Qiu *et al*. introduced DCCA to multimodal emotion recognition [12], and Liu *et al*. examined the robustness of DCCA [3].

Let $X_1 \in \mathbb{R}^{N \times d_1}$ be the instance matrix for the first modality of DCCA and $X_2 \in \mathbb{R}^{N \times d_2}$ be the instance matrix for the second modality of DCCA. Here, N is the number of instances, and d_1 and d_2 are the dimensions of extracted features for these two modalities, respectively.

We have two neural networks that compute representations of multiple modalities by passing them through multiple stacked layers of nonlinear transformation.

Let us use $f_1(X_1)$ and $f_2(X_2)$ to represent the network outputs. The weights, W_1 and W_2, of these networks are trained through standard backpropagation to maximize the CCA objective:

$$(u_1, u_2, W_1, W_2) = \underset{W_1; W_2}{\operatorname{argmax}} \operatorname{corr}\ u_1^\top f_1(X_1), u_2^\top f_2(X_2). \tag{4}$$

After training the neural networks, Qiu *et al*. proposed a weighted sum fusion method [12]:

$$O = \alpha_1 f_1(X_1) + \alpha_2 f_2(X_2), \tag{5}$$

where α_1 and α_2 are parameters that are set by the user. The fused features are used to train a classifier to recognize different emotions. The best output dimensions and the optimal fusion coefficients α_1 and α_2 must be searched by using the prior knowledge of the user and considerable experimentation.

The main shortcoming of DCCA is that only two input modalities are allowed. In addition, users need to set the fusion parameters α_1 and α_2 by experience.

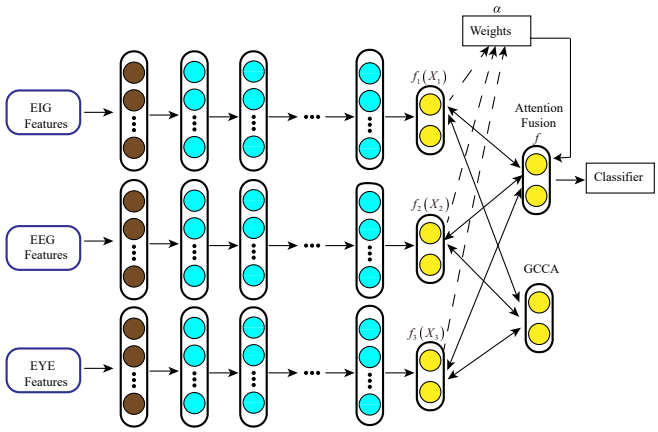


Fig. 1. The framework of DGCCA-AM. It consists of deep networks, GCCA constraint, attention weights, feature fusion layer, and classifier. The GCCA parameters are updated to maximize the GCCA constraints of different modalities. At the same time, the related parameters are updated to minimize classifier loss.

III. Methods

A. Feature extraction

For the SEED-V dataset, differential entropy (DE) features are extracted from EEG signals using a short-time Fourier transform (STFT) with 4 s nonoverlapping Hanning window [5] [6]. These features are divided into five frequency bands: δ (1-4 Hz), θ (4-8 Hz), α (8-14 Hz), β (14-31 Hz), and γ (31-50 Hz). In this way, at every time step, we have DE features of 62 channels, each of which contains data in 5 frequency bands. Finally, the DE features extracted from EEG contain 310 dimensions in total.

As for eye movement features, the same method used in [7] is applied to extracting thirty-three-dimensional features, including pupil diameter, dispersion, and so on. The detailed features are listed in Table I.

TABLE I
THE DETAILS OF FEATURES EXTRACTED FROM EYE MOVEMENT SIGNALS.

Eye movement parameters	Output dimensions
Pupil diameter (X and Y)	Mean, standard deviation and PSD (or DE) in four bands: 0-0.2 Hz, 0.2-0.4 Hz, 0.4-0.6 Hz, and 0.6-1 Hz
Dispersion (X and Y)	Mean, standard deviation
Blink duration (ms)	Mean, standard deviation
Saccade	Mean, standard deviation of saccade duration (ms) and saccade amplitude (°)
Event statistics	Blink frequency, fixation frequency, fixation duration maximum, fixation dispersion total, fixation dispersion maximum, saccade frequency, saccade duration average, saccade amplitude average, saccade latency average

As for eye images, Guo *et al*. proposed an efficient deep neural network model for combining CNN and LSTM networks to extract high-level features. They applied two deep residual networks (ResNet) pretrained on ImageNet to reduce the dimensions of both left and right eye images, two LSTM layers to extract features, and a fully connected layer as an output layer. After feature extraction, EIG features contain 512 dimensions [8].

B. DGCCA-AM

In this section, we introduce DGCCA and the attention mechanism for multimodal emotion recognition. DGCCA extends DCCA from two modalities to arbitrarily numerous modalities. In training, DGCCA passes the input vectors in each modality through multiple layers of nonlinear transformation and backpropagates the gradient of the GCCA objective with respect to network parameters to tune each of the modal networks. To fuse multiple modalities and recognize emotions, we use the attention mechanism for modal fusion and Softmax classifier for classification. While the model uses the gradient of GCCA to adjust the network, the classifier also backpropagates the gradient to tune related networks. Fig. 1 depicts the framework of our proposed model.

1) Deep generalized canonical correlation analysis: Let $(X_1, X_2, \quad , X_i, \quad , X_J)$ denote the input modalities and $X_i \in \mathbb{R}^{d_i \times N}$ be the instance matrix for the i^{th} modality of signals for $i = 1, 2, \quad , J$. Here, N is the number of instances, and d_i represents the dimensions of extracted features. We assume that the network of the i^{th} modality has K_i layers. The k^{th} layer in the i^{th} modality has C_k^i units. The output layer has o units, and the output of the k^{th} layer for the i^{th} modality is as follows:

$$h_k^i = s(W_k^i h_{k-1}^i + b_k^i), \tag{6}$$

where $s : \mathbb{R}^{c_k^i} \to \mathbb{R}^{c_k^i}$ is a nonlinear activation function, $W_k^i \in \mathbb{R}^{c_k^i \times c_{k-1}^i}$ is a matrix of weights and $b_k^i \in \mathbb{R}^{c_k^i}$ is a vector of bias. We denote the output of the final layer as $f_i(X_i) \in \mathbb{R}^{o \times N}$.

The goal of DGCCA is to learn parameters $W^i = \{W_1^i, \ldots, W_K^i\}$ and $b_i = \{b_1^i, \ldots, b_K^i\}$ by solving the following optimization problem:

$$\begin{aligned} \underset{U_i \in \mathbb{R}^{o \times r}; G \in \mathbb{R}^{r \times N}}{\text{minimize}} & \sum_{i=1}^{J} \left\| G - U_i^\top f_i(X_i) \right\|_F^2 \\ \text{subject to } & GG^\top = I_r. \end{aligned} \tag{7}$$

where $f_i(X_i)$ represents the network outputs of the i^{th} modality, and U_i represents a linear transformation of the i^{th} modality.

We solve the DGCCA optimization problem using stochastic gradient descent and adopt backpropagation to update the weights matrices W_i and bias vectors b_i. We now show a sketch of the gradient derivation. The detailed derivation is given in [4]. The solution to the optimization problem is transformed to an eigenvalue problem.

In particular, we define a scaled empirical covariance matrix of the i^{th} network output as $C_{ii} = f(X_i) f(X_i)^\top \in \mathbb{R}^{o \times o}$. Here, $P_i = f(X_i)^\top C_{ii}^{-1} f(X_i) \in \mathbb{R}^{N \times N}$ is the corresponding projection matrix. It is easy to determine that P_i is symmetric and idempotent. We define $P_s = \sum_{i=1}^{J} P_i$. Since P_s is the sum of positive semidefinite P_i, P_s is also positive semidefinite.

Obviously, we can see that the rows of G are the top r (orthonormal) eigenvectors of P_s, and $U_i = C_{ii}^{-1} f(X_i) G^\top$. Then, we can rewrite the objective function as follows:

$$\begin{aligned} & \sum_{i=1}^{J} \| G - U_i^\top f_i(X_i) \|_F^2 \\ = & \sum_{i=1}^{J} \left\| G - G f_i(X_i)^\top C_{ii}^{-1} f_i(X_i) \right\|_F^2, \\ = & rJ - \mathrm{Tr}(GMG^\top). \end{aligned} \tag{8}$$

As indicated in Eq. (8), to minimize the GCCA objective, it is formulated as maximizing $\mathrm{Tr}\left[GMG^\top\right]$, which is the sum of eigenvalues $L = \sum_{i=1}^{r} \lambda_i(M)$.

By taking the derivative of L with respect to each output layer $f_i(X_i)$, we have the following:

$$\frac{\partial L}{\partial f_i(X_i)} = 2U_i G - 2U_i U_i^\top f_i(X_i). \tag{9}$$

Therefore, the gradient is the difference between the r-dimensional auxiliary representation G embedded into the subspace spanned by the columns of U_i (the first term) and the projection of the actual data in $f_i(X_i)$ onto the subspace mentioned above (the second term). Intuitively, if the auxiliary representation G is far away from the modal-specific representation $U_i^\top f_i(X_i)$, then the network weights should receive a large update. The time complexity of computing the gradient descent is $O(JNrd)$, where $d = \max(d_1, d_2, \ldots, d_J)$ indicates the largest dimensions of the input modalities.

2) Attention-mechanism-based feature fusion: Attentive neural networks have recently demonstrated success in a wide range of tasks ranging from question answering to machine translation and image captioning. In this section, we propose an attention-mechanism-based feature fusion method for multimodal emotion recognition [9].

Let $F_j \in \mathbb{R}^{o \times J}$ be a matrix consisting of the j^{th} instance of each output layer $[f_1^j(X_1), f_2^j(X_2), \ldots, f_J^j(X_J)]$, where $f_i(X_i)$ is the result in the output layer of the i^{th} modality, J is the number of modalities, and $f_i^j(X_i) \in \mathbb{R}^o$. The joint representations of all j^{th} instances are formed by the weighted sum of the vectors in F_j:

$$\beta = tanh(F_j), \tag{10}$$

$$\alpha = softmax\left(w^T \beta\right), \tag{11}$$

$$r_j = F_j \alpha^T, \tag{12}$$

where $w \in \mathbb{R}^J$ is the trained parameter vector, and w^T is the transpose of w. The dimensions of α and r are J and o, respectively. We then obtain the final attention mechanism based fusion representations:

$$f^j = tanh(r_j). \tag{13}$$

In this model, we use a Softmax classifier to predict the label $\hat{y}$ for the fusion extracted features. Let X denote the input modalities. The classifier takes the fusion representations as inputs:

$$\hat{p}(y|X) = \mathrm{softmax}\left(W^{(X)} f^j + b^{(X)}\right), \tag{14}$$

$$\hat{y} = \arg\max_y \hat{p}(y|X). \tag{15}$$

The cost function is the negative log-likelihood of the true class label:

$$L = -\frac{1}{M} \sum_{j=1}^{M} t_j \log(y_j) + \lambda \|\theta\|_F^2, \tag{16}$$

where M is the number of instances, and λ is an L_2 regularization hyperparameter.

Now, we have two gradients: one is from DGCCA reconstruction, and the other is from the softmax classifier. We set different learning rates $r_{DGCCA} = 0.01$ and $r_{Classifier} = 0.001$ and update gradients at the same time.

IV. Experiment settings and results

A. The SEED-V Dataset

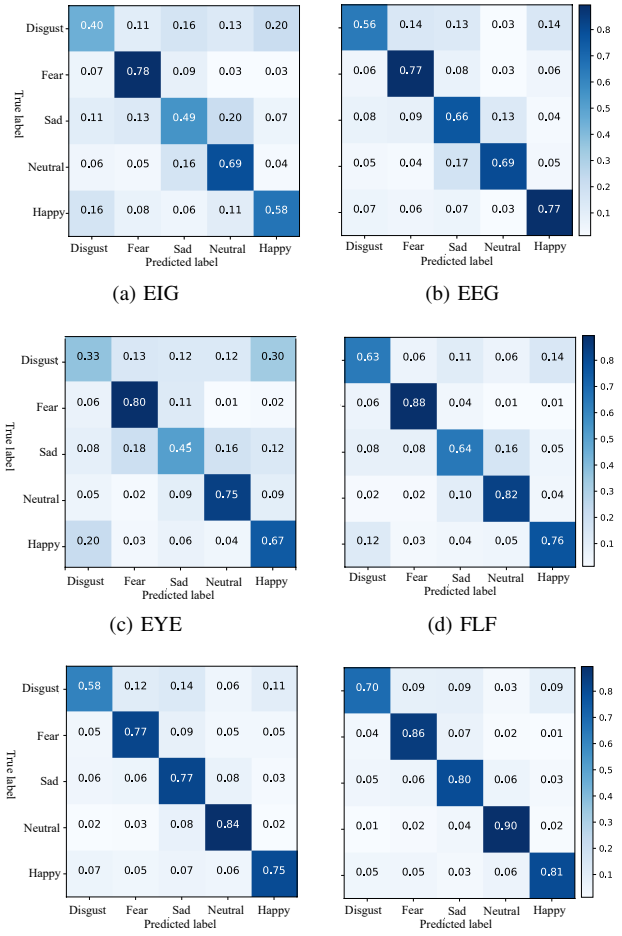
We evaluate our model with respect to the SEED-V dataset. The EEG, eye movement and eye image data of participants were recorded by EEG cap and eye tracking glasses simultaneously.

The SEED-V dataset contains five emotions: happy, sad, fear, disgust, and neutral [14], [15]. Sixteen healthy subjects (10 males and 6 females) aged from 19 to 28 years are selected in total. For each emotion, 9 emotional movie clips are chosen according to the ratings of the elicitation effect reported by subjects after watching clips. The durations of the video clips range from two to four minutes. The experiments contain 3 sessions, each of which consists of 15 randomly played clips.

The EEG signals, containing 62 channels, are recorded with an ESI NeuroScan System at a sampling rate of 1000 Hz. During the experiment, eye movement signals are recorded simultaneously with SMI ETG eye tracking glasses.

Three-fold cross-validation is adopted so that the 15 segments in each session are equally split into three parts. For the convenience of subsequent three-fold cross-validation, each fold is guaranteed to contain 5 clips with different labels.

B. Experimental Results



(a) EIG (b) EEG

(c) EYE (d) FLF

(e) DGCCA-AWS (f) DGCCA-AM

Fig. 2. Confusion matrices of EIG, EEG, EYE, FLF, DGCCA-AWS and DGCCA-AM for five-category emotion recognition on the SEED-V dataset. (a), (b), (c) and (d) are confusion matrices from [8]. Each row of the confusion matrices represents the target class, and each column represents the predicted class. The element (i, j) is the percentage of samples in class i that is classified as class j.

TABLE II
SUMMARY OF DGCCA PARAMETERS FOR SEED-V DATASET

Modalities	Hidden layers	Hidden units	Output dimensions
EEG	3	100±20,30±20	20±10
EYE	3	100±20,30±20	20±10
EIG	4	300±100,100±20,30±20	20±10

The parameters of the model are described in Table II. We select the best combination of parameters and guarantee that the output dimensions of different modalities are the same for each subject. The output dimension is subject-dependent, and each person has a characteristic optimal output dimension.

TABLE III
AVERAGE ACCURACY (%) AND STANDARD DEVIATION (%) OF DIFFERENT FEATURES AND METHODS IN SEED-V. THE TOP 5 ROWS ARE RESULTS FROM [8].

Modality		Fusion method	
EEG	Avg	None	68.22
EYE	Avg	None	62.73
EIG	Avg	None	60.72
ALL	Avg	FLF	73.96
	Std		10.94
ALL	Avg	BDAE	79.63
	Std		6.93
ALL	Avg	DGCCA-AWS	74.66
	Std		5.87
ALL	Avg	DGCCA-AM	**82.11**
	Std		**2.76**

Table III lists the results of some typical existing methods and DGCCA on the SEED-V dataset. Guo *et al*. applied feature level fusion (FLF) and bimodal deep autoEncoder (BDAE) on SEED-V to fuse modalities and considered SVM with linear kernel as a classifier [8]. We verify the single DGCCA with average weighted sum (DGCCA-AWS) and DGCCA-AM on SEED-V. As can be observed, the accuracy of the multimodality for each type of emotion recognition is much higher than for a single modality. This phenomenon indicates that the three modalities have strongly complementary characteristics for five emotions. From Table III, we find that attention weight fusion is effective and achieved the best performance, i.e., accuracy of 82.11% and standard deviation of 2.76%.

Figure 2 depicts the respective confusion matrices of EIG, EEG, EYE, FLF, DGCCA-AWS and DGCCA-AM for five-category emotion recognition on the SEED-V dataset. The element (i, j) is the percentage of samples in class i that is classified as class j. By comparing Fig. 2(f) with the other subfigures in Fig. 2, DGCCA-AM largely improves the classification performance for each of the five emotions, which means that the DGCCA-AM model offers higher fusion efficiency than other methods. By comparing Fig. 2(e) and Fig. 2(f), DGCCA-AM performs better with respect to five-category emotion recognition than does DGCCA-AWS. This result demonstrates that attention mechanism fusion is more effective in fusing modalities than average weighted sum.

Fig. 3 shows the average attention weight distributions of all subjects with respect to the test dataset. We take the average of the attention weights for all subjects to reduce noise interference. From Fig. 3, the test instances of the same emotion from different trials exhibit similar attention weight distributions. This means that for the same emotion, the contributions of EEG, EYE, and EIG to emotion recognition are stable and exhibit small fluctuations with the change of trial. From Fig. 3,

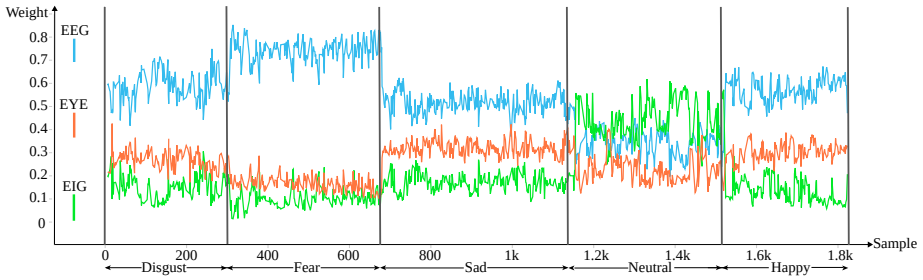


Fig. 3. Illustration of average attention weights for test samples. The ordinate is the attention weights of the three modalities, and the abscissa is the test samples. The weight of each emotion results from the concatenation of the corresponding weights of multiple video clips at different times. Blue, orange and green lines represent weights for EEG, EYE and EIG signals, respectively.

for disgust, fear, sad and happy emotions, we see that EEG contributes most to emotion recognition, followed by EYE, and that EIG contributes the least. This phenomenon indicates that EEG contains more emotion-related information and that EYE and EIG contain less information. It also proves that physiological signals present a greater advantage than other signals in emotion recognition. By comparing Fig. 2 and Table III, we find that the greater the average classification accuracy of a single modality, the greater the corresponding attention weight. However, in comparing Fig. 2 and Fig. 3, the attention weights do not always exhibit the same change pattern as the classification accuracy of the single modality for individual emotions. This might be because the single modality only utilizes the characteristics of this modality, rather than considering the interaction between the different modalities, and this leads to a certain noncorrespondence between the classification accuracy and the corresponding weight for each of the five different emotions.

V. CONCLUSIONS

In this paper, we have introduced DGCCA to five-category emotion recognition. We have proposed a new DGCCA with emotion-related attention mechanism for feature fusion. We have evaluated the performance of our proposed DGCCA-AM on the SEED-V dataset and compared it with the existing approaches. The experimental results demonstrate that DGCCA-AM is superior to the existing methods and achieves the state-of-the-art performance.

VI. ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Program of China (2017YFB1002501), the National Natural Science Foundation of China (61673266 and 61976135), SJTU Trans-med Awards Research (WF540162605), the Fundamental Research Funds for the Central Universities, and the 111 Project.

REFERENCES

[1] H. Hotelling, "Relations between two sets of variates." *Breakthroughs in Statistics*, Springer, pp. 162-190, 1992.

[2] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," *ICML'13*, pp. 1247–1255, 2013.

[3] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using deep canonical correlation," *arXiv preprint arXiv:1908.05349*, 2019.

[4] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, R. Arora, "Deep generalized canonical correlation analysis," *arXiv preprint arXiv:1702.02519*, 2017.

[5] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *IEEE NER'13*, pp. 81–84, 2013.

[6] L.-C. Shi, Y.-Y. Jiao, and B.-L. Lu, "Differential entropy feature for EEG-based vigilance estimation," in *IEEE EMBS'13*, pp. 6627-6630, 2013.

[7] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and EEG to enhance emotion recognition," in *IJCAI'15*, vol. 15, pp. 1170–1176, 2015.

[8] J.-J. Guo, R. Zhou, L.-M. Zhao, and B.-L. Lu, "Multimodal emotion recognition from eye image, eye movement and eeg using deep neural networks," in *IEEE EMBS'19*, pp. 3071-3074, 2019.

[9] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *ACL'13*, vol. 2, pp. 207-212, 2013.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, pp. 32-80, 2001.

[11] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 38, pp. 1798-1828, 2013.

[12] J.-L. Qiu, W. Liu, and B.-L Lu, "Multi-view emotion recognition using deep canonical correlation analysis," in *ICONIP'18*, pp. 221-231, 2018.

[13] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, "Multimodal emotion recognition using EEG and eye tracking data," in *IEEE EMBS'14*, pp. 5040–5043, 2014.

[14] L.-M. Zhao, R. Li, W.-L. Zheng, and B.-L. Lu, "Classification of five emotions from eeg and eye movement signals: Complementary representation properties," in *IEEE NER'19*, pp. 611–614, 2019.

[15] T.-H. Li, W. Liu, W.-L. Zheng, and B.-L. Lu, "Classification of five emotions from eeg and eye movement signals: Discrimination ability and stability over time," in *IEEE NER'19*, pp. 607–610, 2019.

[16] H. Leng, Y. Lin, and L. Zanzi, "An experimental study on physiological parameters toward driver emotion recognition," in *EHAWC'07*. Springer, pp. 237–246, 2007.

[17] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[18] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[19] J. Kim and E. Andr´e, "Emotion recognition based on physiological changes in music listening," *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 30, pp. 2067–2083, 2008.

[20] M. L.-H. Võ, A. M. Jacobs, L. Kuchinke, M. Hofmann, M. Conrad, A. Schacht, and F. Hutzler, "The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect," *Psychophysiology*, vol. 45, no. 1, pp. 130–140, 2008.

[21] Y. Wang, M. Huang, X. Zhu, and L. Zhao, " Attention-based LSTM for aspect-level sentiment classification," in *EMNLP'16*, pp. 606-615, 2016.