Individual-Specific Emotion Recognition from Multimodal Physiological Signals

Paper ID: 1333

Anonymous Author(s)

ABSTRACT

Emotion recognition from multimodal physiological signals has attracted universal research efforts in recent years. The majority of current studies approach the problem through establishing universal correlations among physiological signals and emotion states after integrating multimodal information at the feature or decision level. However, these methods do not consider individual differences or capture complementary correlations. To address this problem, we propose Individual-specific Hypergraph Neural Networks (I-HGNN), which take into account individual differences in model construction. Every type of physiological signals is used to establish an independent hypergraph with the hyperedges depicting the correlations among vertices (subject, stimuli). Specifically, we have designed an individual embedding hypergraph to formulate individual differences in physiology and psychology. Correlation information mined in a single modality and the influence of individual differences can be projected into the subspace of other modalities through correlation-complementary fusion mechanisms self-adaptively. In doing so, we transform the emotion recognition task as classification problem of the vertices in a multi-hypergraph framework. Experimental results and comparisons with the state-of-the-art methods on the DEAP and ASCERTAIN datasets demonstrate the superiority of the proposed method.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → **Supervised learning by classification**; **Neural networks**; • **Applied computing** → **Bioinformatics**.

KEYWORDS

Emotion recognition, Physiological Signal, Individual difference

ACM Reference Format:

Anonymous Author(s). 2019. Individual-Specific Emotion Recognition from Multimodal Physiological Signals: Paper ID: 1333. In *MM '19: ACM International Conference on Multimedia,October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/1122445.1122456

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '19, October 21–25, 2019, Nice, France

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

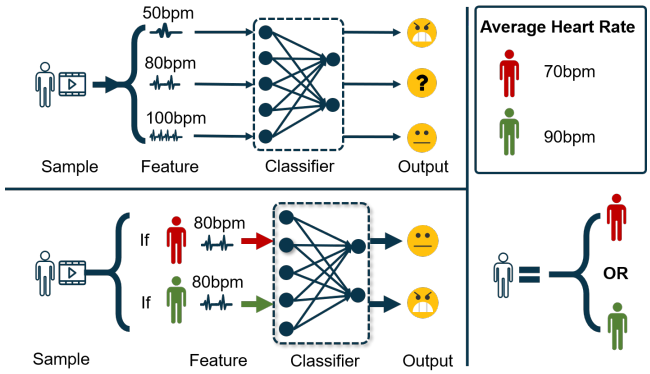


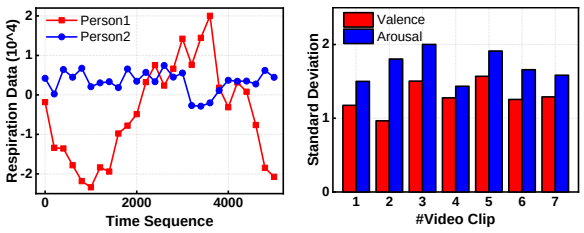
Figure 1: Illustration of traditional emotion recognition methods and the method weighting individual difference. Suppose that the heart rate of a person will rise when he/she is exciting and decline when calm.

1 INTRODUCTION

Emotion recognition is aimed at narrowing the communicative gap between the humankind and the computer. Computational systems are developed to recognize human emotions based on bodily and physiological signals with accompanied emotional changes[22]. With the rapid development of social media and internet of things, the demands for analyzing the emotional data are growing fast. Under such circumstances, improving emotion recognition has become an urgent issue and also a fundamental challenge in human-computer interaction. According to the type of modality input, existing emotion recognition models can be divided into models based on language[32], facial expressions[22], body language[3, 15], and physiological signals[soleymani2012multimodal]. As an objective and instantaneous channel for emotion expression[calvo2010affect], physiological signals have an unique advantage in many emotion recognition applications like driver monitoring systems[20].

Generally, emotion recognition can be considered as a classification problem. However, traditional classification methods cannot be directly applied in emotion recognition based on physiological signals due to their following specific characteristics: (1) physiological signals recorded by different types of sensors independently are of heterogeneity and multiple modalities[27]; (2) autonomic response patterns are a complex non-linear function of both the stimulus and the subject[6]; and (3) due to the high requirements for conditions of data collection, only hundreds of samples are available in existing released datasets[17, 27].

As noted in [4], the majority of existing emotion recognition methods from physiological signals focus on tackling the first challenge. Typically, a two-step pipeline is employed, i.e. feature extraction and sample classification. For the collected multimodal physiological signals, these methods mainly aim to extract discriminative and robust features from each modality [25] and then combine them in the classification step by some fusion strategies[16, 22]. However, there are several issues with above-mentioned methods on the affective gap challenge:



(a) (b)

Figure 2: Left:the RES change of two subjects with the same emotional response when watching the same video. Right: The valence and arousal standard deviations of the 32 subjects on the several video clips in DEAP dataset.

First, these methods mainly employ feature-level-fusion (FLF) and decision-level-fusion (DLF) to integrate information from multimodal physiological signals, which cannot well deal with the missing data or overlook the interaction among modalities[1]. FLF simply concatenates feature vectors into a massive vector or matrix, which is vulnerable to the imperfectness of data and curse of dimensionality. In contrast, DLF focuses on the combination the outputs of base learners in a more efficient way, which enhances the robustness but still overlooks the low-level or correlation-level interaction among different modalities[1].

Second, these methods usually train a classifier to establish patterns universally applicable to all subjects (stimulus-response specificity), without considering the difference among subjects in terms of physiological and psychological features (individual-response specificity). As shown in Figure 1, suppose that we have two subjects, subject A and subject B and their average heart rate. Given a sample whose rate is 70-90, classfiers will hardly make any precise prediction merely based on the information without knowledge of specific physiological information of the very owner of the sample. In addition to physiological side, the difference in psychological features can also be captured by two constructs: valence/arousal focus, demonstrating the extent to which subjects attend the hedonic/arousal component of their emotional experience[7].

Third, the CNN-based methods [28] with limited training samples tend to cause overfitting and reduce the robustness of the model. For example, the DEAP dataset[17], a relatively large dataset currently available, includes data of physiological signals of 32 volunteers watching 40 video clips. Moreover, the quality of the data of available emotion datasets are not properly guaranteed, which could also lead to overfitting.

In this paper, we propose a novel fusion mechanism at the correlation-level and also resolves the data insufficiency problem. Specifically, we present a novel modelIndividual-Specific Hypergraph Neural Networks for ER into which correlation-complementary fusion mechanisms and individual embedding hypergraph are introduced. In the first place, one hypergraph is established for one modality with the same vertices(subject, stimuli) a dnhyperedges depicting the associations among physiological signals generated under various stimuli. In doing so, we convert emotion recognition into classification of vertices in the multi-hypergraph. As for individual embedding hypergraph, we calculate the emotional tendency of every subject to formulate individual difference in psychological attributes. Subsequently, the vertices of the same subject under similar stimuli are connected by hyperedges to strengthen the links among these vertices. Second, the impact of multi-modal hypergraph are acquired automatically by the network. Thus, relative correlation mined in different modalities and the influence of individual difference could be projected into the subspace of other modalities according to the relevance matrix. Moreover, we use a fully-connected network for combination of the embedded features from different hypergraph neural networks. Within this framework, we can construct a model weighting the influence of individual difference in emotion recognition, explore complex inter-subject correlations and integrate information of multiple physiological signals. We have tested the effects of our method with DEAP dataset and ASCERTAIN dataset. Results and comparison with the state-of-the-art methods have proved the superiority of the proposed method.

The main contributions of this paper are three-fold:

- We have designed a novel correlation-level fusion method. Our method calculates the correlation coefficient of each modality adaptively and avoids the excess of feature dimensons. The efficiency of the interaction of discriminative information is therefore improved.
- Inspired by individual-response specificity, we propose the working framework of I-HGNN which first quantifies individual difference in physiological and psychological attributesthrough hypergraph. Moreover, the consideration of the influence of individual difference improves the precison and personalization level of emotion recognition through adding correlation-level fusion into the mode.
- A number of experiments have been conducted with the DEAP and ASCERTAIN dataset, concluding that the proposed method, I-HGNN, is superior to the state-of-the-art methods and can tackle the problem of data missing.

2 RELATED WORK

This section gives a brief review of existing works concerning ER from physiological signals, multimodal fusion methods and individual-response specificity.

2.1 Emotion recognition from physiological signals

Compared to other emotion recognition approach, identification of physiological signals based on a multimodal framework is receiving wider and wider attention because of its objectivity and

instantaneity. In [14], a distributed wireless system is developed to assess stress resistance during stressful training to measure HRV , EEG, GSR. Electromyography (EMG) simultaneously and skin conductance is employed to determine emotion in real time to address affective gaming in [21]. [26] investigated eye movement patterns of subjects when they watched clips of a variety of emotions to verify the assumption that emotional content draws eye fixations and strengthens memory for the scene gist while weakening the assumption of encoding of peripheral scene details. User-centered implicit affective indexing employing emotion detection based on physiological signals, such as electrocardiography (ECG), galvanic skin response (GSR), electroencephalogra-phy (EEG) and face tracking, has achieved significant progress through applying a quality adaptive multimodal fusion scheme[11]. In addition to purely physiological signals, RECOLA[23] provided an emotional corpus including audios, video clips, ECG and EDA signals of 46 subjects recorded simultaneously. [12] assessed personal response with multi-modal emotional and physiological signals including Electroencephalogram (EEG), Electrocardiogram (ECG), Galvanic Skin Response (GSR) .

2.2 Multimodal Fusion Methods

For the natural process of emotional expression of humans,one kind of physiological signals does not suffice to depict the change of physiological state caused by emotional change. Different types of physiological signals can capture complementary information âĂŤ something invisible in an individual modality itself. Therefore, the majority of currently available datasets and emotion recognition studies are based on multimodal framework.

[28] proposed to extract features of different modalities as an independent vector and concatenates different vectors into a new two-dimensional matrix as the input of CNN. However, feature-level fusion is plagued by several limitations. Hughes [13] has proven that an increase in the feature set may cause decline of classification accuracy if the the size of the training set is insufficient. Takahashi [30] used FLF of EEG signals and peripheral physiological signals but failed to improve classification accuracy. Integrating results of several classifiers operating on different modalities corresponding to a separate feature space can make up for the deficiency. [9] utilized EEG and eye gaze data as inputs to obtain the posterior probabilities of each modality, and achieve confidence measure summation fusion. Compared to single-modality classification, the simple fusion method significantly improves its proven performance for emotion recognition. An ensemble combination strategy was proposed to train a classifier with the best overall performance according to the ranking of classification accuracy of different modality classifiers[16]. As summed up by [1],DLF overlooks low-level interaction among modalities.

2.3 Individual-Response Specificity

The majority of current models depicting the internal emotional response system of human body can be categorized into two types, stimulus-response specificity and individual-response specificity. Most exiting emotion computation studies are focused on the generic correlations among physiological signals and emotions, while overlooking individual difference in physiological[31] and psychological [7] state in responding to stimuli. However, giving weight to the difference in computation can enhance accuracy and analysis of the features of an individual in responding to different stimuli will make the result more personalized and accurate. As shown in Figure 2, there is significant differences among subjects in terms of the changing trend and features of physiological signals after receiving the same video stimuli. The emotions different subjects perceive on the same video clips are also different. To further explore the intra-and inter-individual variability, EMG signals of medium-distance male runnersare collected in the same conditions[10], little variation was found for each muscle of a subject in terms of peak time across trials and EMG profile, but there were significant difference among subjects. ($P < 0.01$). [12] investigated significant individual difference in alpha peak frequency of EEG signal among subjects.[2][30]made a significant contribution to understanding of the structure of emotional experience. Following the two studies, many were devoted to[29][19]the difference in individual emotional experience. For instance, [27] argue that emotion recognition is influenced by individual differences in terms of the scope and dicrepanycy of personnal emotional experience.

3 METHOD

In this section, we will give a detailed introduction to our I-HGNN framework. Our target is to predict the emotional state from multimodal physiological signals of the samples considering the individual differences (subject,video stimulus).

Given samples represented by vertices, we construct one hypergraph for each modality. The input of the vertices in the hypergraph represent the features extracted from corresponding modalities. Following that, we input the emotional tendency of the subjects into vertices and use hyperedges to connect vertices constructed with the same subject that receives similar stimuli so as to establish the individual hypergraph. The hyperedge features of the modality are obtained through edge convolution in every hypergraph and are fed into the correlation argument fusion block. Every modality will generate a hyperedge feature reinforced with information from other modalities. Subsequently, the node features acquired through hypergraph convolution operation independently conducted in every modality are incorporated together through an FC layer to obtain emotions predicted by the node. Figure 3 illustrates the detailed flowchart of our framework.

3.1 Hypergraph Construction

To process the features generated by physiological signals of different modalities, we construct an unique hypergraph for each modality. A vertex in a hypergraph represents a composition of a single subject and a stimulus received by the subject, while hyperedges are employed to establish connections among the vertices in the hypergraph. This method will be pointless if all the vertices are the same for the m hypergraphs (the value of m depends on the type of physiological signals). Suppose that the training set $S = S_1, S_2, ..., S_N$ with features $\mathbf{X}^{(i)} = \mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, ..., \mathbf{x}_N^{(i)}$ from modality i, with S_j denoting the j-th training sample and vector $\mathbf{x}_j^{(i)}$ denoting the feature of the j-th training sample from modality

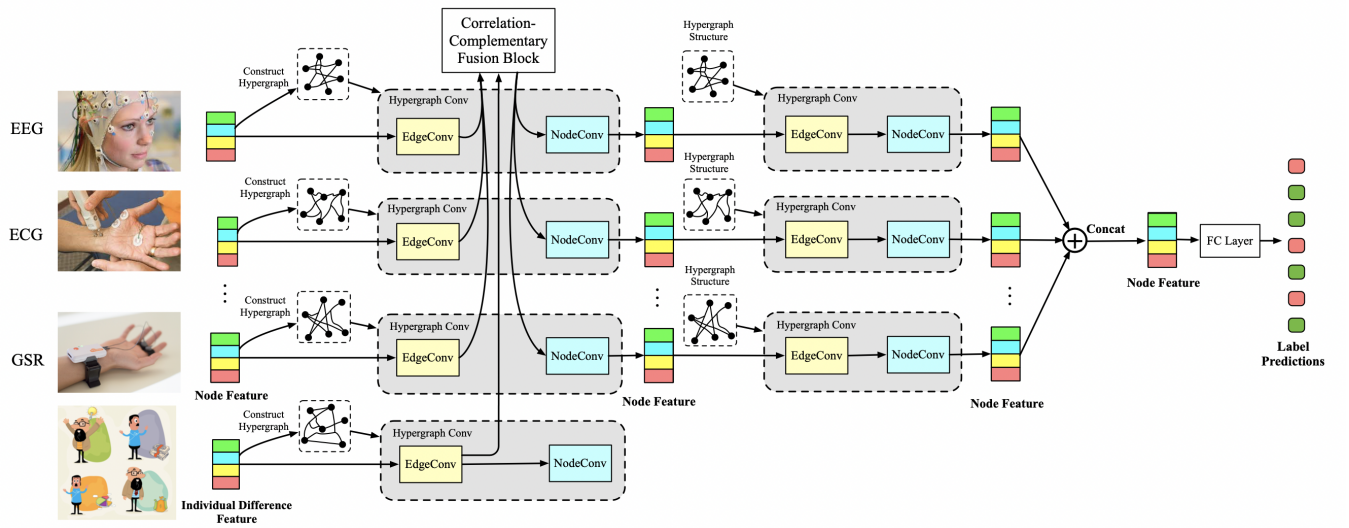


Figure 3: Our I-HGNN framework is composed of two types of hypergraphs:individual hypergraph and physiological-signal-based hypergraph; two important sub-structures:hypergraph convolution and correlation-complementary fusion block. Hypergraph convolution includes two kinds of operations, EdgeConv and NodeConv.

i. A hypergraph $\mathcal{G}^{(i)} = \{\mathbf{V}^{(i)}, \mathbf{E}^{(i)}, \mathbf{W}^{(i)}\}$ can be built, with $\mathbf{V}^{(i)}$ representing the set of vertices, $\mathbf{E}^{(i)}$ representing the set of hyperedges and $\mathbf{W}^{(i)}$ representing a weight matrix for hyperedges. Hyperedges are generated by the kNN method in our study. Specifically, each vertex will act as the centroid to build hyperedges once and only once. Hyperedge e_a is centered on vertex v_a and the k nearest vertices to v_a will share the hyperedge. The distance measure between two vertices implemented in our work is the Euclidean distance between corresponding feature vectors. With hyperedges generated in the process, the incident matrix $\mathbf{H}^{(i)}$ is employed to represent hypergraph structure in hypergraph neural network. The correlation between vertex a and vertex b is denoted by the matrix element $h_{a,b}$. We can model it as follows:

$$h_{a,b}^{(i)} = \begin{cases} exp\left(-\dfrac{d(\mathbf{x}_a^{(i)}, \mathbf{x}_b^{(i)})^2}{d^2}\right) & b \in \mathbf{u}_a, \\ 0 & b \notin \mathbf{u}_a, \end{cases} \tag{1}$$

where $d(\mathbf{x}_a^{(i)}, \mathbf{x}_b^{(i)})$ denotes the Euclidean distance of the a-th sample and b-th sample in feature space. Because of deficiency of prior knowledge regarding the significance of hyperedges, the weight matrix $\mathbf{W}^{(i)}$ is set to be completely the same matrixes in our model. As a result, the incident matrix $\mathbf{H}^{(i)}$ will contain the complete information of the entire hypergraph.

3.2 Individual Hypergraph Construction

The strength and type of emotional response are determined by the intensity of the stimulus and certain systems of human body. Existing computational models ignore the individual differences in physiological and psychological perspectives when constructing models, thus leading to the deviation of the models from reality. Inspired by the theory of [6][7][10], we model the emotional response system of human body in the physiological and psychological dimension.

Following[7], we attempt to assign weights to individual differences by computing the emotional tendency of every subject in the two dimension of Valence and Arousal, respectively named as the valence focus and the arousal focus. Taking the valence for example, all samples (s subjects $*t$ video clips) are divided into the training set $\mathbf{T_r}$ and the testing set $\mathbf{T_e}$ to construct an emotion matrix $\mathbf{V} \in \mathbb{R}^{s \times t}$:

$$V_{a,b} = \begin{cases} +1, & (a,b) \in T_r \quad \&\& \quad (a,b) = HV, \\ 0, & (a,b) \in T_e, \\ -1, & (a,b) \in T_r \quad \&\& \quad (a,b) = LV, \end{cases} \tag{2}$$

where (a, b) represents the sample corresponding to subject a and video clip b.

Within the framework, the vector of emotional tendency of subjects $f_v \in \mathbb{R}^{s \times 1}$ can be computed by:

$$f_v = V \cdot \mathbf{1}_{t \times 1}, \tag{3}$$

where $f_v(a)$ represents valence focus of subjects a.

The emotional tendency of each subject is added into the model as features of the corresponding vertices of an individual hypergraph. Individual difference in physiological attributes also plays a significant role. For example, the body temperature of a person in a cheerful mood rises temporarily. Suppose there are a group of subjects, the average daily temperature thereof is 36.5°C. The temperature of a chosen subject at the moment is 36.5°C. Traditional models would reach the conclusion that the chosen subject is in a stable mood. However,the average daily temperature of subject a is 37°C and the average daily temperature of subject b 36°C. If the subject a is chosen, we will predict that the subject is in high spirits; On the contrary, if subject b is chosen, we will predict that

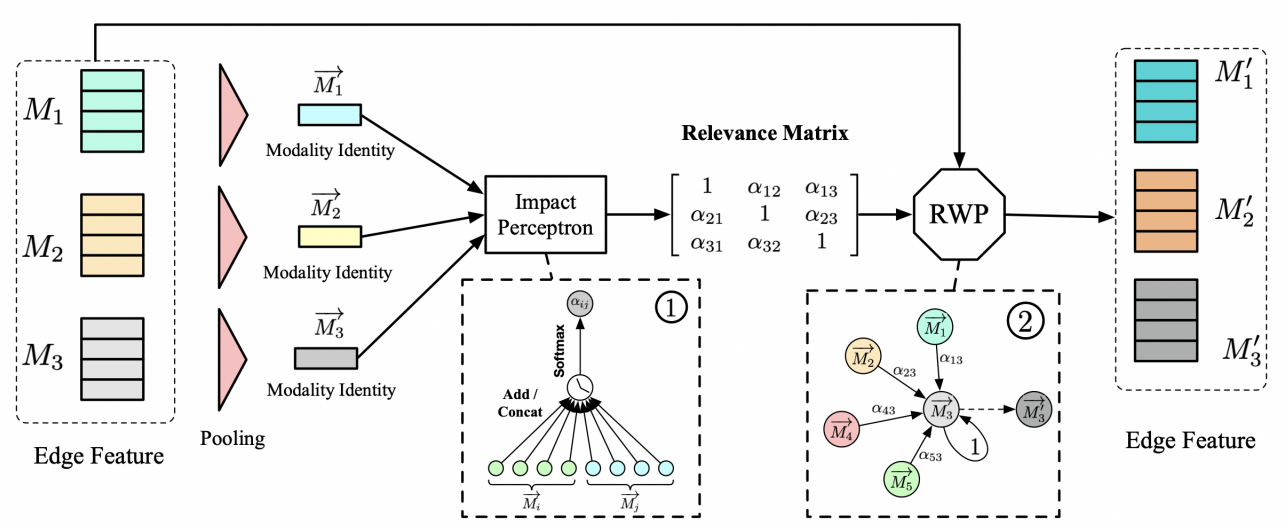


Figure 4: Correlation-complementary fusion block. It adopts edge feature matrixes generated by EdgeConvin multiple hypergraphs as input, acquires inter-modality relevance coefficients through impact impact perceptron. RWP means relevance weighted pooling, every modality realize edge feature fusion with other modalities according to the relevance matrix.

the subject is in low spirits. Therefore, traditional models might generate inaccurate predictions in such conditions. In contrast, our emotion recognition model attaches importance to such individual difference in physiological and psychological response to emotional change. Correspondingly, our emotion computation model assigns more weight to the discrepancies among physiological signals of an unique subject under similar stimuli. This can be achieved through establishing hyperedges among vertices of a subject.

To measure the extent of similarity among video clips, we compute the data of each video clips as following:

$$c_v = V \cdot \mathbf{1}_{1\times s}, \tag{4}$$

where $c_v(b)$ represents average scores of video clips b in Valence.

Therefore, we integrate the vertices of an individual as a set in establishing hyperedges. Inside a set, kNN is used to establish hyperedges. Each vertex is chosen as a centroid for once and the hyperedges are determined after all vertices have served as the centroid. Note that the framework is of high flexibility. For some special contexts of use, other measurement indicators and connecting patterns of hyperedges can be adopted to represent individual difference in physiological and psychological features.

After the establishment of individual hypergraph, hypergraphs of different modalities interact with one another through correlation-complementary fusion mechanism presented in the following part. The emotional tendency of subjects will be projected into the hypergraph of other modalities for diffusion. The correlations among similar vertices in hypergraphs based on multimodal physiological signals will be reinforced to varying degrees. The degree can be understood as the intensity of the influence of individual difference. The next section will present details of the process.

3.3 Hypergraph Convolution

For each modality, we adopt hypergraph convolution operation [8] for node representation. For the vertices feature $\mathbf{X} \in \mathbb{R}^{N\times C_1}$ in a single modality, we can obtain a embedded vertices feature representation Y after applying hypergraph convolution operation. In hypergraph convolution process, compared with node features, hyperedges are more useful as to reflect the nature of each modal. Thus, we split hypergraph convolution into two separate components: EdgeConv and NodeConv, which is illustrated in Figure 3. EdgeConv operation is formulated by:

$$\mathbf{M} = \mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^\top\mathbf{X}\mathbf{\Theta}, \tag{5}$$

where $\mathbf{W} = \mathrm{diag}(\mathbf{w_1}, \ldots, \mathbf{w_e})$ is weight factors for each hyperedge. $\mathbf{\Theta} \in \mathbb{R}^{C_1\times C_2}$ is a trainable parameter in EdgeConv. $\mathbf{H}$ denotes hypergraph incidence matrix. $\mathbf{D}_e$ is the diagonal matrices of hyperedge degrees. $\mathbf{M} \in \mathbb{R}^{\mathbf{E}\times\mathbf{C_2}}$ is a intermediate product in hypergraph convolution operation called hyperedge feature. NodeConv is defined as:

$$\mathbf{Y} = \mathbf{D_v^{-1}}\mathbf{H}\mathbf{M}, \tag{6}$$

where $\mathbf{D}_v$ denotes the diagonal matrices of node degrees. Y is the output of the NodeConv, which can be used for classification.

Combining Eq. (5) and (6), we can obtain the final hypergraph convolution formulation in this work:

$$\mathbf{Y} = \mathbf{D}_v^{-1}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^\top\mathbf{X}\mathbf{\Theta}, \tag{7}$$

Then we can build a hypergraph convolutional layer in the following formulations:

$$\begin{cases} \mathbf{M}_{(l+1)} = \sigma(\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^\top\mathbf{X}_{(l)}\mathbf{\Theta}_{(l)}) \\ \mathbf{X}_{(l+1)} = \sigma(\mathbf{D}_v^{-1}\mathbf{H}\mathbf{M}_{(l+1)}), \end{cases} \tag{8}$$

where σ is the activation function, $\mathbf{X}_{(l+1)}$ represents the hypergraph signal at l layer. e

3.4 Correlation-Complementary Fusion

Inter-sample correlations in hypergraphs are depicted with hyperedges. Thus, modalities in our model also interact through edge feature matrix. Feature matrices of the modalities $\mathbf{M_i} \in \mathbb{R}^{N\times c}$ are acquired through edge convolution in the convolutional layer, in which c is a hyperparemeter in the model. However, a more global representation form of features needs to be mined to better represent characteristics of modalities. Thus, a new $1 \times c$ hypergraph factor m is generated in each modality through max-pooling.

Impact Perceptron: to measure inter-modality influence and optimize information interaction among modalities, we concatenate the hyper factor m_i and m_j from modility i, j and input the result into the perceptional layer to acquire $\alpha_{i,j}$ self-adaptively. The value of α represents the intensity of the interaction. Moreover, to avoid the explosive growth of output, we also add softmax and Relu functions into the model, which can normalize the output to range $[0, 1]$. With results of ergodic combinations, we set heritability rate of the modalities themselves as 1 and gain a relevance matrix with a diagonal line of 1.

Relevance Weighted Pooling: discriminative correlation information mined in every modality is not the same. Effectively integrating the information can improve robustness of the model. As is shown in Figure 4, the edge feature matrix of every modality reinforces other modalities by their weights in the relevance matrices. In the end, newly acquired reinforced edge feature matrices are applied in point convolution of each modality.

Original correlations mined in various modalities through correlation-level fusion can realize effective interaction. The influence of individual difference could be diffused into other modalities self-adaptively, thus improving the accuracy of emotion recognition. After being processed by the two layers of graph convolution networks, node features generated in various modalities are concatenated and input into the FC layer to generate the result of emotion prediction.

4 EXPERIMENT SETUP

4.1 Dataset

DEAP is the most widely recognized and used multimodal physiological signal dataset for human emotion analysis. The dataset collects 32 participants' electroencephalogram (EEG) and peripheral physiological signals that were generated as they watched 40 1-minute clips. Subjects rated their own change on valence and arousal level when watching every video clips. In the end, self-evaluation with a discrete scale of 1-9 was used as ground truth. Following settings in [28], we use 5 as the threshold and divide rating results into "high" and "low".

ASCERTAIN is another public multimodal physiological signals dataset of video clips response. 58 student volunteers (mean age = 30, 21 female) participated in the study. The subjects watched 36 clips lasting for 51-127 seconds. Furthermore, the subjects rated their emotional experience with seven-point scale of valence(very positive to very negative) and arousal (very exciting to very boring). Physiological signals that included ECG, GSR, Frontal EEG were precisely collected simultaneously. Note that there is missing data to varying degrees for each type of signals due to body movement and experimental equipment limitations. For instance, maximum missing data is noted for GSR for the sensitivity of the GSR device to body movement.

Different feature extracting methods are employed in the two datasets due to the difference in data formats and modalities. We extract features of physiological signals, including EEG, EMG and RES, of DEAP dataset following the settings in [25]. We extract the same features from ASERTAIN dataset that includes GSR, EEG, and ECG[27].

4.2 Baselines

To compare our the method with the state-of-the-art ones for ER, we select the following methods as baselines: (1) DNN and CNN[28], (2) Bayesian Classifier (BC)[5], (3) Segment Level Decision Fusion (SLDF)[24] , (4) the naive Bayes (NB) and linear SVM classifiers [27] and (5) the NB and SVM using the decision fusion technique (NB-DF and SVM-DF)[18].

Note that the formats and collection of the physiological signals of modalities in ASCERTAIN dataset that we have selected are all distinct from DEAP dataset, which can further prove the robustness of our model. However, methods merely based on DEAP dataset cannot be migrated to ASERTAIN dataset to make comparisons.

4.3 Implementation Details

In implementation, I-HGNN uses a two-layer hypergraph neural network to process each modality. The dimension of the hidden layer is set to 8. We choose ReLu as the activating function for hypergraph convolution layers and set dropout rate of both layers at 0.5 to avoid overfitting. During the pre-training and fine-tune stage, the cross-entropy loss is used as criterion and Adam optimizer with a learning rate of 0.001 is employed. The percentage of training data ia set at 90%, and the remaining 10% was used for testing. We perform 10 runs and reported the averages of the results to avoid the influence of randomness. For a fair comparison, we carefully tune the parameters of the baselines and report the best results.

5 RESULTS AND ANALYSIS

5.1 Comparison with the State-of-the-art

First, we compare the performance of the proposed method with the state-of-the-art approaches in DEAP dataset and ASCERTAIN. The experiment results and comparisons are shown in Table 1. In experiments, we have compared the proposed I-HGNN with various models using different fusion approaches. As shown in Table 1, our proposed method can achieve the best performance on DEAP dataset with the classification accuracy of 84.36% in valence and 76.75% in arousal. Compared with CNN using physiological signals of five modalities, our I-HGNN has gains of 3.6% in valence and 4.6% in arousal. As shown in Table 2, our proposed method can achieve the best performance in ASCERTAIN dataset with the classification accuracy of 85.43% in valence and 73.52% in arousal. Compared with NB-DF using physiological signals of four modalities, our I-HGNN has gains of 22.9% in valence and 5.7% in arousal.

The better performance of the proposed method can be attributed to the following reasons:

1) For a certain modality, hypergraph structure can mine complex high-order correlations among samples in feature space. In most

Methods	Fusion	Modality	Accuracy%	
			Valence	Arousal
DNN	FLF	EEG,GSR,etc.	75.78	73.13
CNN	FLF	EEG,GSR,etc.	81.41	73.36
BC		EEG	66.6	66.6
SLDF	DLF	EEG	76.9	69.1
I-HGNN-I	CLF	EEG,GSR,ECG	81.68	74.48
I-HGNN-IF		EEG,GSR,ECG	78.32	73.66
I-HGNN	CLF	EEG,GSR,ECG	**84.36**	**76.75**

Table 1: Compared results in DEAP dataset (CLF stands for Correlation-Level Fusion)

Methods	Fusion	Modality	Accuracy%	
			Valence	Arousal
SVM		EEG,GSR,etc.	64.47	65.38
NB		EEG,GSR,etc.	65.46	68.39
SVM-DF	DLF	EEG,GSR,etc.	66.32	67.6
NB-DF	DLF	EEG,GSR,etc.	69.48	69.53
I-HGNN-I	CLF	EEG,EMG,RES	83.74	72.87
I-HGNN-IF		EEG,EMG,RES	80.34	72.18
I-HGNN	CLF	EEG,EMG,RES	**85.43**	**73.52**

Table 2: Compared results in ASCERTAIN dataset (CLF stands for Correlation-Level Fusion)

cases, the correlations are of multiple types. In typical situations, several samples possess a common attribute and the high flexibility of hyperedges makes it suitable for description of latent correlations.

2) Muitimodal physiological signals have three features. First, the correlations among physiological signals of different modalities are multi-dimensiona. Second, the total number of feature dimensions may be too great, increasing the difficulty of computation. Third, missing data is a common problem. Currently available fusion methods like FLF and DLF cannot effectively deal with multimodal data featured by such physiological singals. Different from feature-level and decision-level fusion, the method proposed in the study conducts correlation-level fusion on different modalities with hyperedge feature matrixes. In this way, the correlation among different modalities will be computed through networks in a self-adaptive manner, and therefore higher-level interaction of discriminative information can be achieved.
3) Present methods are focused on establishing generic functional relations and do not take into consideration the personal feature of a subject. Individual hypergraph quantifies the emotional tendency of subjects and intensifies interaction among signal samples of a subject under similar stimuli. In the end, the influence of individual difference will be projected into feature space of other modalities to varying degrees through correlation-level fusion and therefore enhances the accuracy of emotion recognition for every subject.

5.2 Ablation Study

Correlation-complementary fusion mechanism and individual hypergraph proposed in I-HGNN plays a vital role. In this sub-section, we further evaluate the influence of the two structures on the recognition performance through using different combinations of the components. The results are shown in Figure 5. In the chart,"I-HGNN-IF" represents both two components being removed from our mode. "I-HGNN-I" represents only deleting the remaining model of individual hypergraph. It can be clearly seen that in the two datasets, removing individual hypergraph will diminish the performance significantly. Moreover, if correlation-level fusion is further removed, the model is converted into multi-modal decision-level fusion, with a further decline of precision.

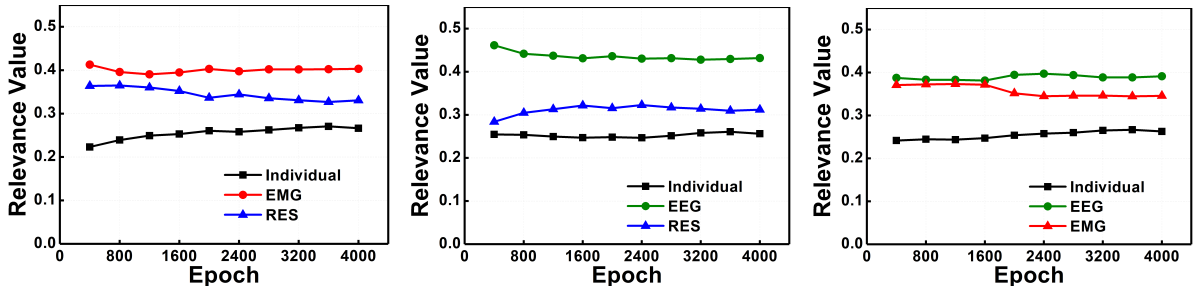
5.3 On Parameter Sensitivity

In our method, the establishment of hypergraphs is a critical step. An essential hyperparameter k in this process decides the smoothness of the hypergraph structure. To assess the influence of the selected neighbor number k in hyperedge generation on the performance of our method, we test the model with varying value of k. When $K = 1$, the hypergraph is simplified into a simple graph. Figure 6 provides the recognition accuracy performance curve with respect to the variation of parameter k on the DEAP dataset and the ASCERTAIN dataset, respectively.

It can be clearly seen that the performance on the two datasets is significantly stable with k in a wide range. Take valence of DEAP dataset for example. When $k < 8$, the performance will improve with the increase of the value of k due to the fact that when k is too small, the high-order relationship among different vertices cannot be fully mined. This proves the effectiveness of our method in modelling complex correlations among data. However, the performance slowly decline when k is greater than 8. Because each hyperedge connects too many vertices. DEAP dataset only includes 32 subjects and 40 video clips. In the hypergraph generated with physiological signals, mismatching physiological signals are connected by the same hyperedges. In an individual hypergraph, physiological response stimulated by video clips of different types are incorrectly connected, possibly limiting the discriminative ability of the hypergraph structure. Thus, the performance will decline quickly if K is set at a too great value. We reach the conclusion that both a too small and too great value of K will cause degeneration of representation ability and thus degrading of the performance of the model.

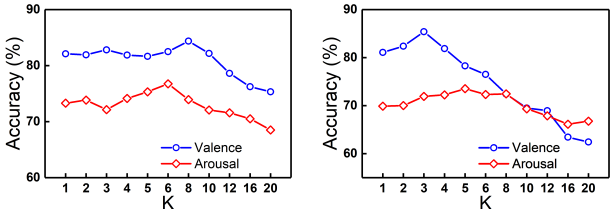
5.4 On Relevance Matrix Convergence

Every single modality strengthens its discriminative information through integrating hyperedge feature matrixes of other modalities with correlation-argument fusion mechanisms. In this process, the strength of the influence of other modalities on the modality is not the same. It is acquired by the network in a self-adaptive manner. The values in the general relevance matrix represent the importance



(a) The relative impact of other modalities on EEG (b) The relative impact of other modalities on EMG (c) The relative impact of other modalities on RES

Figure 5: The change of Multimodal relevance coefficient by epochs.



(a) DEAP dataset (b) ASCERTAIN dataset

Figure 6: The influence of hyperedge generation parameter K on emotion recognition performance.

of the interaction among modalities. The convergence of the matrix is the foundation for the stability of fusion.

To explore the convergence pattern of the relevance matrix in model training and the intensity of interaction among modalities, we have used DEAP dataset to predict the change of the relevance matrix by epoch in the Arousal process as an example. As is shown in Figure 5, every graph represents the changing trend of the intensity of the influence of other modalities on a certain modality. The black line representing individual difference tends to become stable. The correlation information contained in this modality will be rapidly embedded in other modalities. Another interesting result is that the EEG signal shows significant reinforcing effects probably due to the strong correlations between the EEG signal and emotional change. On the whole, the relevance matrix shows good convergence, which is beneficial for model stability.

6 CONCLUSION

The paper proposes an integrating framework of multimodal hypergraph neural network for emotion recognition based on physiological signals. To the best of our knowledge, it is the first time that individual difference in psychological and physiological features is quantified and added into a model to make more personalized prediction of emotions. The approach takes into consideration complex high-order correlations among samples of a single modality and the interinfluence among modalities at the correlation level. Every compound vertex consists of a subject and a stimulus and the correlations among samples are depicted by hyperedges. Individual difference in physiological and psychological features is represented by the features of vertices and hyperedges of the individual hypergraph. The complementary information of an individual modality and the influence of individual difference are projected into feature space of other modalities in a self-adaptive manner through correlation-complementary mechanisms. In the end, features and the importance of different modalities are jointly explored through a fully-connected network. We have conducted experiments on two public datasets and results show that the proposed method sinificantly outperforms other recent methods. The proposed model is also of high flexibility in incorporating new information in the multi-hypergraph structure.

REFERENCES

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443.

[2] Lisa Feldman Barrett, James Gross, Tamlin Conner Christensen, and Michael Benvenuto. 2001. Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition & Emotion* 15, 6 (2001), 713–724.

[3] Daniel Bernhardt and Peter Robinson. 2007. Detecting affect from non-stylised body motions. In *International conference on affective computing and intelligent interaction*. Springer, 59–70.

[4] Rafael A Calvo and Sidney D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing* 1, 1 (2010), 18–37.

[5] Seong Youb Chung and Hyun Joong Yoon. 2012. Affective classification using Bayesian classifier and supervised learning. In *2012 12th International Conference on Control, Automation and Systems*. IEEE, 1768–1771.

[6] Bernard T Engel. 1960. Stimulus-response and individual-response specificity. *AMA Archives of General Psychiatry* 2, 3 (1960), 305–313.

[7] Lisa A Feldman. 1995. Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of personality and social psychology* 69, 1 (1995), 153.

[8] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2018. Hypergraph Neural Networks. *arXiv preprint arXiv:1809.09401* (2018).

[9] Kiel Gilleade, Alan Dix, and Jen Allanson. 2005. Affective videogames and modes of affective gaming: assist me, challenge me, emote me. *DiGRA 2005: Changing Views–Worlds in Play.* (2005).

[10] Laura Guidetti, Gianfranco Rivellini, and Francesco Figura. 1996. EMG patterns during running: Intra-and inter-individual variability. *Journal of Electromyography and Kinesiology* 6, 1 (1996), 37–48.

[11] Rishabh Gupta, Mojtaba Khomami Abadi, Jesús Alejandro Cárdenes Cabré, Fabio Morreale, Tiago H Falk, and Nicu Sebe. 2016. A quality adaptive multimodal affect recognition system for user-centric multimedia indexing. In *Proceedings of the 2016 ACM on international conference on multimedia retrieval*. ACM, 317–320.

[12] Saskia Haegens, Helena Cousijn, George Wallis, Paul J Harrison, and Anna C Nobre. 2014. Inter-and intra-individual variability in alpha peak frequency. *Neuroimage* 92 (2014), 46–55.

[13] Gordon Hughes. 1968. On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory* 14, 1 (1968), 55–63.

[14] Emil Jovanov, AO'Donnell Lords, Dejan Raskovic, Paul G Cox, Reza Adhami, and Frank Andrasik. 2003. Stress monitoring using a distributed wireless intelligent sensor system. *IEEE Engineering in Medicine and Biology Magazine* 22, 3 (2003), 49–55.

[15] Asha Kapur, Ajay Kapur, Naznin Virji-Babul, George Tzanetakis, and Peter F Driessen. 2005. Gesture-based affective computing on motion capture data. In *International conference on affective computing and intelligent interaction*. Springer, 1–7.

[16] Jonghwa Kim and Florian Lingenfelser. 2010. Ensemble Approaches to Parametric Decision Fusion for Bimodal Emotion Recognition.. In *BIOSIGNALS*. 460–463.

[17] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2012), 18–31.

[18] Sander Koelstra and Ioannis Patras. 2013. Fusion of facial expressions and EEG for implicit affective tagging. *Image and Vision Computing* 31, 2 (2013), 164–174.

[19] Peter Kuppens. 2008. Individual differences in the relationship between pleasure and arousal. *Journal of Research in Personality* 42, 4 (2008), 1053–1059.

[20] H Leng, Y Lin, and LA Zanzi. 2007. An experimental study on physiological parameters toward driver emotion recognition. In *International Conference on Ergonomics and Health Aspects of Work with Computers*. Springer, 237–246.

[21] Arturo Nakasone, Helmut Prendinger, and Mitsuru Ishizuka. 2005. Emotion recognition from electromyography and skin conductance. In *Proc. of the 5th international workshop on biosignal interpretation*. Citeseer, 219–222.

[22] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[23] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 1–8.

[24] Viktor Rozgić, Shiv N Vitaladevuni, and Rohit Prasad. 2013. Robust EEG emotion classification using segment level decision fusion. In *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 1286–1290.

[25] Mohammad Soleymani, Frank Villaro-Dixon, Thierry Pun, and Guillaume Chanel. 2017. Toolbox for Emotional feAture extraction from Physiological signals (TEAP). *Frontiers in ICT* 4 (2017), 1.

[26] Ramanathan Subramanian, Divya Shankar, Nicu Sebe, and David Melcher. 2014. Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes. *Journal of vision* 14, 3 (2014), 31–31.

[27] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe. 2018. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing* 9, 2 (2018), 147–160.

[28] Samarth Tripathi, Shrinivas Acharya, Ranti Dev Sharma, Sudhanshu Mittal, and Samit Bhattacharya. 2017. Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset.. In *Twenty-Ninth IAAI Conference*.

[29] Michele M Tugade, Barbara L Fredrickson, and Lisa Feldman Barrett. 2004. Psychological resilience and positive emotional granularity: Examining the benefits of positive emotions on coping and health. *Journal of personality* 72, 6 (2004), 1161–1190.

[30] David Watson and Auke Tellegen. 1985. Toward a consensual structure of mood. *Psychological bulletin* 98, 2 (1985), 219.

[31] Chao-Gan Yan, R Cameron Craddock, Xi-Nian Zuo, Yu-Feng Zang, and Michael P Milham. 2013. Standardizing the intrinsic brain: towards robust measurement of inter-individual variation in 1000 functional connectomes. *Neuroimage* 80 (2013), 246–262.

[32] Sicheng Zhao, Hongxun Yao, and Xiaoshuai Sun. 2013. Video classification and recommendation based on affective analysis of viewers. *Neurocomputing* 119 (2013), 101–110.