A Feature Learning Approach Based on Multimodal Human Body Data for Emotion Recognition

H. Perry Fordson, X. Xing, K. Guo and X. Xu

Research Centre of Human Body Data Science, South China University of Technology, GZ, China
eeperryfordson@mail.scut.edu.cn,{xfxing, guokl, xmxu}@scut.edu.cn

Abstract— The human body as an entire structure of a person contains physiological and physical reactions that connect with emotions. Emotions play a crucial role in our day to day activities, not only the way we interact with colleagues but also in our decision making processes. Emotion recognition from multimodal signals allows a direct assessment of the innermost state of a person which is regarded an important component of Human-Computer interactions (HCI). This paper proposes an enhancement feature learning technique using a Hyper Enhanced Learning System (HELS). We conduct an experiment on the DEAP database by utilizing four signals. Signals were preprocessed to remove artifacts and noise and feature extraction is done to obtain relevant features. We then introduced a feature enhancement system that generates random weights and enhances feature nodes. We assorted emotions into three categories and classified the emotional states using Artificial Neural Networks (ANN). The average accuracy rate of our proposed system was 78.6% and 79.9% respectively for valence and arousal. Obtained results show that combining physiological signals is relevant for accurate human emotion recognition task.

I. Introduction

We all experience strong feelings necessary for any living being. These include fear, surprise, joy, happy, anger, and disgust [1]. Emotions are the grassroots of the daily living of a human being and plays a very crucial role in human cognition, namely rational decision making, perception, human interactions, and human intelligence [2, 3] (shown in Figure 1). Moreover, emotions have been widely ignored especially in the field of human-computer interactions (HCI) [4, 5].

Presently, attention has been drawn to its importance and researchers around the world are making efforts aimed at finding better and appropriate ways to uniformly build relationships between the way computers and humans interact. To build a system for HCI, knowledge of emotional states of subjects must be known. Interest in emotion recognition is traditionally from different modalities, for example, facial expressions, body posture, speech, and text [6–9]. These traditional ways are still gaining attention today from scholars even though their reliability and effectiveness may be questioned because they can be deliberately altered. Affective computing is the study and development of systems and devices that can recognize, interpret, process, and simulate human affect [10]. This has emerged to convert technology and emotions in HCI [11, 12] The main design is to simulate emotional interactions between humans and computers by calculating and measuring the emotional states of a subject. A person's inmost emotional state may become evident by subjective experiences (how a person is feeling), internal expressions (physiological and biological signals), and external expressions (audio-visual signals) [13]. In other words, how a person reacts when confronted with an emotional stimulus. Self-assessment reports provide precious information, but generates issues with its validity, certification and corroboration. For example, Mr. Karikari in a flight with his wife during a turbulent may not accurately tell her how he is truly feeling, but, would give an answer to what he expects will make her see him as a strong man.

Emotions are time varying affective phenomena that are elicited as a result of a stimulus. When we are introduced to a particular stimulus, how we respond to it is necessary to access our emotional intelligence [14]. Physiological signals can assist in obtaining a better understanding of a person's response and expression at a time of observation. These involve multiple recordings from both central and nervous systems. Emotional stimuli in short music/videos are introduced to elicit emotions. They are shown to persons in an experimental setting and signals are taken from other parts of their body which enables detecting emotional traces instantaneously. The central nervous system comprises the brain and the spinal cord whiles the autonomous nervous system is a control system that acts unconsciously and regulates bodily functions like heart rate, pupillary response and sexual arousal. Consequently, they can hardly be falsified.

Physiological signals that are spontaneous and highly correlated with human emotion includes electroen-

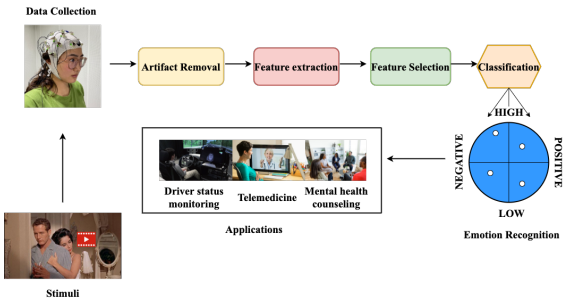

Figure 1. Fundamental Modules for Emotion Recognition

cephalogram (EEG), electromyogram (EMG), galvanic skin response (GSR), heart rate (HR), temperature (T), functional magnetic resonance imaging (fMRI), blood volume pulse (BVP), positron emission topography (PET), and respiration (RES). This is evident in the work of [15] which conducted an extensive review on physiological parameters and their relation to human emotion. In our previous work [16], we used broad learning system (BLS) [17] without the enhancement nodes as a classifier to train the physiological signals for emotion recognition.

In this paper, we aim improve on the construction of the BLS to detect human affective states into valence-arousal dimension using EEG, GSR, EMG, and RES signals. We reported our results in three ways. Two-classes which include positive and negative for valence and high and low for arousal, three-classes using the self reported feedback values which range from 1-9 in valence and arousal axis, and finally, we defined another three classes based on coded emotional keywords. We present our results based on single signal contribution and also combined all signals for multi-modal based classification. We employed the Database for Emotion Analysis using Physiological signals [18]. We applied our enhanced feature learning systems after feature extraction and during feature selection stages and use Artificial Neural Networks (ANN) for classification.

The rest of this paper is structured as follows. Section II presents the proposed method and strides needed for accurate classification results. Experimental procedure including preprocessing, feature extraction, feature enhancement techniques, and ANN classifier are introduced in Section III. Section IV presents how the valence-arousal emotional states are modeled. Section V reports and discusses the obtained results and we conclude our paper in section VI.

II. Methodological Setup

II-A. Hyper Enhanced Learning System

In this section, we introduce the proposed method for emotion recognition. As shown in Figure 2, the system retains but improves the structure of the BLS by replacing feature nodes of BLS with groups of physiologically extracted data to form a hybrid neuro-multimodal network. Also, the BLS takes data directly, we takes the extracted features as inputs to reduce the structure complexity and memory. Lets assume our input feature X, projected using $\Phi_i(XW_{ei}+\beta_{ei})$, is the ith mapped physiological features, F_i, where W_{ei} is randomly generated weights, β_{ei} are bias and Φ is linear transformation. The first i group of mapped physiological features is concatenated by denoting $F^i \equiv [F_1, F_2, ..., F_i]$. Similarly, enhancement feature nodes for the jth group, $\zeta_j(F^iW_{hj}+\beta_{hj})$ is denoted as E_j. The first j group of enhanced nodes are concatenated by denoting $E^j \equiv [E_1, E_2, ..., E_j]$. We then applied linear inverse problem [19] to fine tune the the initial weight, W_{ei}, so as to obtain richer features. Therefore, assuming an input signal X, with N samples each with M dimensions, the output is $Y \in \mathbb{R}^{N\times C}$. For n physiological feature, each mapping randomly yields k nodes which can be represented in the form:

$$F_i = \Phi(XW_{ei}+\beta_{ei}), i = 1, 2, ..., n \tag{1}$$

We denote feature nodes as $F^n \equiv [F_1, ..., F_n]$, and denote the nth group of enhancement node as:

$$E_m \equiv \zeta(F^iW_{hm}+\beta_{hm}) \tag{2}$$

Hence, the hyper enhanced structure, $Y = [F_1, F_2, ..., F_n \mid \zeta(F^iW_{h1} + \beta_{h1}), ..., \zeta_j(F^iW_{hm} + \beta_{hm}]W^m$, i.e., $Y = [F_1, F_2, ..., F_n \mid E_1, E_2, ...E_m]W^m$ is represented as:

$$Y = [F^n \mid E^m]W^m \tag{3}$$

where $W^m = [F^n \mid E^m]^+Y$. The enhancement node is added contemporaneously with the connections of the physiological features. We construct a different structure by linking each group of physiological feature to a group of enhancement node, seen in Figure 3 for multimodal analysis. For an input extracted feature X_N, for n physiological feature and n enhancement group, we have the output enhanced feature $[F_1, \zeta(F_iW_{h1}+\beta_{h1} \mid ...F_n, \zeta(F_nW_{hn}+\beta_{hn})]W^n$. Therefore Y_N

$$\begin{aligned} Y_N = [F_1, ..., F_n \mid \zeta(F_1W_{h1}+\beta_{h1}), ..., \\ \zeta_n(F_nW_{hn}+\beta_{hn})]W^n \end{aligned} \tag{4}$$

where $F_i, = 1, ..., n$, are physiological mapped features gotten from (1). Emotion recognition systems have respective strides that needs to be carefully considered in order to obtain accurate classification results as detailed below. The block diagram of our study is presented in Figure 3.

III. Experiments

III-A. Database For Emotion Analysis Using Physiological Signals (DEAP)

The dataset involves 32 healthy participants who watched a 1 minute, 40 different videos whiles EEG and other peripheral signals including EMG, GSR, and RES are measured. In addition, emotional rating (valence/arousal/dominance/familiarity/liking) in accordance with participants and video ID on a discrete 9-point scale, which qualify the video content or the preference of the experimenter were collected. Table 1 summarizes the content of the DEAP database.

III-B. Data Preprocessing and Feature Learning

In order to obtain robust results in emotion recognition task, data preprocessing, feature extraction and selection, and classification steps are required to be given special attention. Firstly, data were downsampled at 128Hz. For EEG data, electrooculogram (EOG) noise

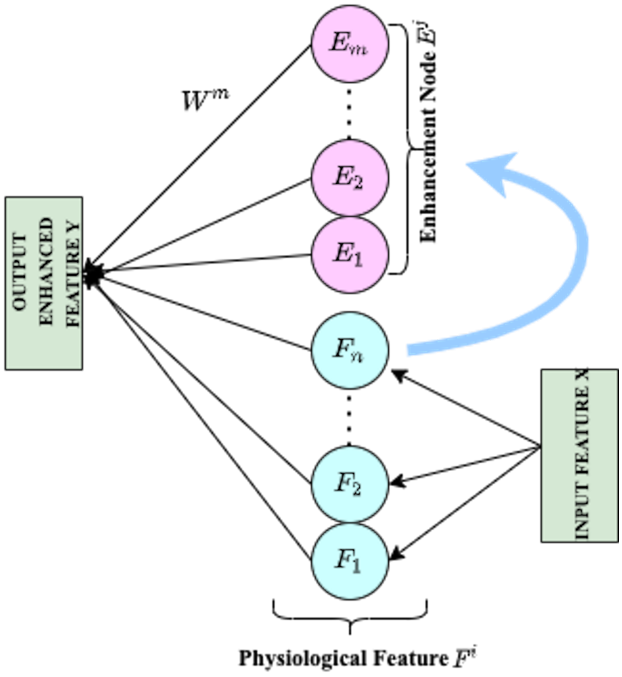

Figure 2. Hyper Enhanced Learning System Construction

Table 1. DEAP Database Summary

Number of participants	32
Recorded signals	EEG, EOG, EMG, GSR, RES, BVP, T
Number of videos	40
Self-report	Valence, arousal, dominance, familiarity, liking, keywords
Ratings scale	1-9

is removed. Then a bandpass frequency filter from 4.0-45.0Hz was applied after which data is averaged to the common reference. EEG data are then segmented into 60 second trials and a 3 second pre-trial baseline is removed. For the EMG, GSR and RES data, the data is also downsampled to 128Hz. Also, the data are segmented into 60 seconds trials and we remove a 3 seconds pre-trial baseline. Following the settings of [11], different features for each physiological signals were extracted. This is shown in Table 2. After feature extraction, we introduce the hyper enhanced learning system to take inputs from each feature of each modality to produced a more informative feature for classification. In the final execution, this study employed a three

Table 2. Feature Extraction for each modality constructed in the following: Video/Trial $\times$ Channel $\times$ Data

Modality	Array Shape	Extracted Features
EEG	$40 \times 32 \times 8064$	Power spectral density in different bands
GSR	$40 \times 1 \times 8064$	Number of peaks, amplitude of peaks, rise time, statistical moments
RES	$40 \times 1 \times 8064$	Main frequency, power spectral density, statistical moments
EMG	$40 \times 2 \times 8064$	Power and Statistical moments

Table 3. Two-Defined Classes in Valence-Arousal Model

Assortment		
Valence	**Arousal**	**Rating (r)**
Negative	High	$r \leq 4.5$
Positive	Low	$4.5 \leq r$

Table 4. Three-Defined Classes in Valence-Arousal Model

Assortment		
Valence	**Arousal**	**Rating (r)**
Unpleasant	Calm	$1 \leq r \leq 3$
Neutral	Average	$4 \leq r \leq 6$
Pleasant	Excited	$7 \leq r \leq 9$

layer artificial neural network to model each modality for single signal classification, and then combined all signals for multimodal classification. We set the dimension of the hidden layer to 16. We choose ReLu activation function and used a dropout rate of 0.5 for all layers to avoid overfitting. We used binary cross-entropy loss as a criterion and Adam optimizer with the 0.001 learning rate. We allocated 80% of our data for training and 20% for testing. The experiment was conducted in a subject-independent setting. We used one subject's data for testing and the rest of the remaining subject's data for training. We repeated results for all subjects and averaged the results.

IV. Valence-Arousal State Modeling

The distinct classes using 1-9 discrete scales within the valence-arousal dimension is presented. This is necessary to find correlation amongst different discrete emotions which correspond to higher levels of a particular emotion. Tables 3 and 4 show modeling of the two and three defined classes. The participants' reported their feelings after watching the affective music video clips. Firstly, for two classes, we assigned "High" and "Low" for arousal and "Positive" and "Negative" for valence. Secondly, the three classes modeling assigned "Calm", "Average", and "Activated" for arousal and "Unpleasant", "Neutral" and "Pleasant" for valence. Finally, we defined valence-arousal using 6 affective coded keywords. These include (1) Happy, (2) Amuse, (3) Sad, (4) Neutral, (5) Surprise, and (6) Angry. This is shown in Table 5.

V. Results and Discussion

This section summarizes and assesses the obtained results for emotion classification in valence-arousal di-

Table 5. Classes in Valence-Arousal Model using Coded Emotional Keywords

Dimension	Affective Classes	Emotion Tagging
Valence	Unpleasant	Angry, Sad
	Neutral	Neutral, Surprise
	Pleasant	Happy, Amuse
Arousal	Calm	Sad, Neutral
	Average	Happy, Amuse
	Activated	Surprise, Angry

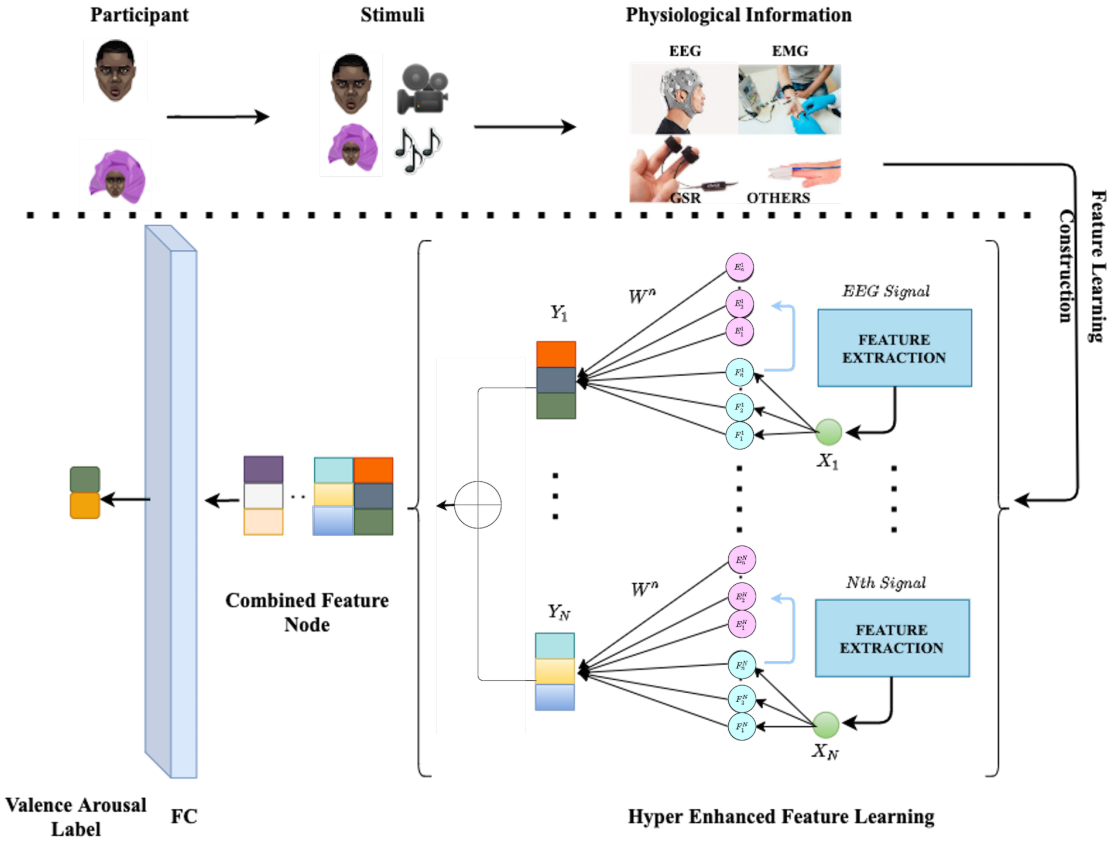

Figure 3. Overall framework of the proposed method using an enhanced feature learning approach to multimodal emotion recognition

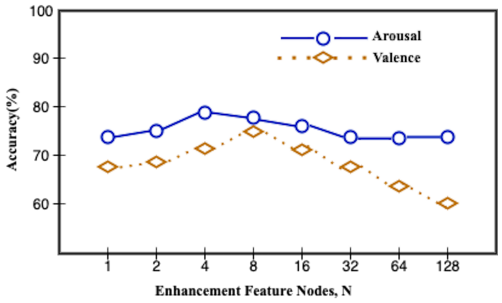

Figure 4. Feature learning influence by enhancement nodes parameter generation, N on classification performance.

Table 6. Classification Accuracy (Two Classes)

Physiological Signal	Accuracy Result % Valence	Arousal
EEG	70.1	72.2
GSR	73.3	75.6
RES	72.8	74.5
EMG	69.8	72.4
EMG+EEG	69.2	71.1
GSR+EEG	71.3	72.2
RES+EEG	68.7	70.2
GSR+EMG	68.8	70.3
RES+EMG	65.9	70.2
RES+GSR	69.4	70.9
GSR+EMG+EEG	71.3	71.4
RES+EMG+EEG	70.8	69.9
RES+GSR+EEG	71.0	71.2
RES+GSR+EMG	71.1	68.9
EEG+GSR+RES+EMG	**78.6**	**79.9**

mension. Our emotional states are presented in two and three defined classes using 1-9 discrete self rating scale and 6 emotional keywords to emphasize the three defined areas in valence arousal dimension.

Tables 6, 7 and 8 presents the classification results for two defined classes, three defined classes, and the emotional keywords in the valence-arousal dimension accordingly.

VI. Conclusions

The rationale behind our proposed multi-modal fusion is such that, GSR for instance is well known to correlate well with arousal scale but poorly with valence. Thus, using different multi-modal signal combinations separately for arousal and valence may improve the classification accuracy. Initially, we classified each physiological signal to obtain a single modal classification result

Table 7. Classification Accuracy (Three Classes)

Physiological Signal	Accuracy Result %	
	Valence	Arousal
EEG	67.4	69.6
GSR	67.3	70.2
RES	68.4	69.7
EMG	66.6	68.0
EMG+EEG	66.2	68.5
GSR+EEG	67.4	70.1
RES+EEG	65.2	67.4
GSR+EMG	67.5	66.4
RES+EMG	67.9	65.3
RES+GSR	68.4	69.1
GSR+EMG+EEG	68.4	70.1
RES+EMG+EEG	67.1	69.8
RES+GSR+EEG	68.2	70.0
RES+GSR+EMG	69.1	69.9
EEG+GSR+RES+EMG	**69.7**	**71.9**

Table 8. Classification Accuracy (Coded Emotional Keywords)

Physiological Signal	Accuracy Result %	
	Valence	Arousal
EEG	68.1	70.6
GSR	69.2	72.9
RES	69.7	71.1
EMG	68.5	70.8
EMG+EEG	67.3	69.1
GSR+EEG	68.6	70.6
RES+EEG	66.2	68.1
GSR+EMG	67.9	68.4
RES+EMG	68.4	67.8
RES+GSR	69.2	66.6
GSR+EMG+EEG	69.2	71.3
RES+EMG+EEG	68.9	70.5
RES+GSR+EEG	69.0	71.1
RES+GSR+EMG	70.2	71.6
EEG+GSR+RES+EMG	**72.2**	**75.4**

and find which signal best classifies human emotion. Then, we fused all signals to obtain a multimodal fusion. After, we compared our results to related works.

The performance of the DEAP dataset is notably steady with N, in a broad range clearly seen in Figure 4. In the arousal dimension, when $N \in [1,4]$, the performance improves as there is an increase in the value of N, the enhancement feature nodes. The observation is based on the fact that, performance increases as N increases. When N is small, the high relationship between features cannot be fully mined. Also, the figure shows a decline in performance when N is greater than 5 and then stability in performance thereafter. In contrast, the valence dimension increases in performance when $N \in [1,8]$. Similar to the arousal dimension, performance increases as there is an increase in enhancement nodes. The figure also shows that as N is greater than 8, the performance declines in the valence dimension.

In Tables 9 and 10, we compared our obtained results with recently published works in both two and three classes for emotion recognition task in the valence-arousal dimension. Our results prove to be promising

Table 9. Experimental comparison with related work in two and three classes.

Dimension	Two Classes			Three Classes		
				1-9 values		6 Coded Keywords
	Ours	[18]	[11]	Ours	[16]	Ours
Valence %	**78.6**	57.0	69.6	**69.7**	61.3	**72.2**
Arousal %	**79.9**	52.3	70.1	**71.9**	60.0	**75.4**

Table 10. Comparison with related work

Works	Valence %	Arousal %
[18]	62.7	57.0
[20]	57.0	52.3
[11]	69.6	70.1
[3]	78.0	74.0
Ours	**78.6**	**79.9**

and more robust. This is indicative of the fact that, the proposed enhancement learning system generates enhanced feature nodes that are more significantly informative than those chosen in earlier studies. In this paper, we present hyper enhanced learning system, the proposed feature learning approach to multimodal emotion recognition using physiological signals. Given a physiological signal information, we preprocess data by removing artifacts and noise to make it smooth. After, several features are extracted. These features are then mapped as inputs to construct an enhanced hybrid neuro-multimodal learning network that automatically updates weights with enhancement nodes to generate more informative features nodes. The model then learns complex relationships within signals and explore the importance of different modalities through a fully connected neural network. We used the DEAP database for our experiment and by comparison with other works, we show the supremacy of the proposed method. We established two and three class modeling in the valence-arousal dimension using discrete rating values from 1-9. We also use 6 emotionally coded keywords to define the three areas in the valence-arousal dimension. Results were reported for single and multimodal signals. Fusions of multimodal signals prove to be more robust than using a single modality for emotion recognition.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant U1801262 and Grant 61972163, in part by Key-Area Research and Development Program of Guangdong Province, China, under grant 2019B010154003, in part by the Natural Science Foundation of Guangdong Province, China, under Grant 2018A030313474, Grant 2020A1515010781, and Grant 2019B010154003, in part by the Guangzhou Key Laboratory of Body Data Science, under Grant 201605030011, in part by Science and Technology Project of Zhongshan, under Grant 2019AG024, and in part by the Fundamental Research Funds for the Central Universities, SCUT, under Grant 2019PY21, and Grant 2019MS028.

REFERENCES

[1] P. Ekman, "An Argument for Basic Emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[2] E. L. Johnson, J. W. Kam, A. Tzovara, and R. T. Knight, "Insights into human cognition from intracranial EEG: A review of audition, memory, internal cognition, and causality," p. 051001, 2020.

[3] Y. Luo, Q. Fu, J. Xie, Y. Qin, G. Wu, J. Liu, F. Jiang, Y. Cao, and X. Ding, "EEG-Based Emotion Classification Using Spiking Neural Networks," *IEEE Access*, vol. 8, pp. 46 007–46 016, 2020.

[4] F. Ren and Y. Bao, "A review on human-computer interaction and intelligent robots," *International Journal of Information Technology and Decision Making*, vol. 19, no. 1, pp. 5–47, 2020.

[5] Y. Yun, D. Ma, and M. Yang, "Human–computer interaction-based Decision Support System with Applications in Data Mining," *Future Generation Computer Systems*, vol. 114, pp. 285–289, 2021.

[6] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou, and J. Mao, "A facial expression emotion recognition based human-robot interaction system," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 668–676, 2017.

[7] C. L. Reed, E. J. Moody, K. Mgrublian, S. Assaad, A. Schey, and D. N. McIntosh, "Body Matters in Emotion: Restricted Body Movement and Posture Affect Expression and Recognition of Status-Related Emotions," *Frontiers in Psychology*, vol. 11, p. 1961, 2020.

[8] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[9] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowledge and Information Systems*, vol. 62, no. 8, pp. 2937–2987, 2020.

[10] Y. Li, R. Kumar, W. S. Lasecki, and O. Hilliges, "Artificial intelligence for HCI: A modern approach," *Conference on Human Factors in Computing Systems - Proceedings*, 2020, pp. 1–8.

[11] X. Zhang, J. Liu, J. Shen, S. Li, K. Hou, B. Hu, J. Gao, T. Zhang, and B. Hu, "Emotion Recognition From Multimodal Physiological Signals Using a Regularized Deep Fusion of Kernel Machine," *IEEE Transactions on Cybernetics*, pp. 1–14, 2020.

[12] S. Wang, J. Li, T. Cao, H. Wang, P. Tu, and Y. Li, "Dance Emotion Recognition Based on Laban Motion Analysis Using Convolutional Neural Network and Long Short-Term Memory," *IEEE Access*, vol. 8, pp. 124 928–124 938, 2020.

[13] P. Mishra and N. Salankar, "Automation of Emotion Quadrant Identification by Using Second Order Difference Plots and Support Vector Machines," *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2020, pp. 1–4.

[14] L. Hajncl and D. Vučenović, "Effects of measures of emotional intelligence on the relationship between emotional intelligence and transformational leadership," *Psihologijske Teme*, vol. 29, no. 1, pp. 119–134, 2020.

[15] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors (Switzerland)*, vol. 18, no. 7, p. 2074, 2018.

[16] Perry Fordson and X. Xu, "Research on emotion recognition and feature learning method based on Multimodal human data," *Dissertation, South China University of Technology. https://cdmd.cnki.com.cn/Article/CDMD-10561-10118875306.htm.*, pp. 1–53, 2018.

[17] C. L. Chen and Z. Liu, "Broad Learning System: An Effective and Efficient Incremental Learning System Without the Need for Deep Architecture," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 1, pp. 10–24, 2018.

[18] S. Koelstra, C. Mühl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A Database for Emotion Analysis Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[19] T. Goldstein, B. O'Donoghue, S. Setzep, and R. Baraniuk, "Fast alternating direction optimization methods," *SIAM Journal on Imaging Sciences*, vol. 7, no. 3, pp. 1588–1623, 2014.

[20] M. Soleymani, F. Villaro-Dixon, T. Pun, and G. Chanel, "Toolbox for Emotional feAture extraction from Physiological signals (TEAP)," *Frontiers in ICT*, vol. 4, p. 1, 2017.