IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 22, 2019, accepted April 30, 2019, date of publication May 3, 2019, date of current version May 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2914872

Exploiting EEG Signals and Audiovisual Feature Fusion for Video Emotion Recognition

BAIXI XING[1,2], HUI ZHANG[1], KEJUN ZHANG[1], LEKAI ZHANG[1], XINDA WU[2], XIAOYING SHI[2], SHANGHAI YU[2], AND SANYUAN ZHANG[1]
[1]College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
[2]School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China

Corresponding author: Kejun Zhang (zhangkejun@zju.edu.cn)

This work was partly supported by the Natural Science Foundation of Zhejiang Province of China under Grant LY19F020047 and Grant LZ19F020002, the National Key Research and Development Program of China under Grant 2017YFB1002600, and the National Natural Science Foundation of China under Grant 61402141.

ABSTRACT External stimulation, mood swing, and physiological arousal are closely related and induced by each other. The exploration of internal relations between these three aspects is interesting and significant. Currently, video is the most popular multimedia stimuli that can express rich emotional semantics by its visual and auditory features. Apart from the video features, human electroencephalography (EEG) features can provide useful information for video emotion recognition, as they are the direct and instant authentic feedback on human perception with individuality. In this paper, we collected a total of 39 participants' EEG data induced by watching emotional video clips and built a fusion dataset of EEG and video features. Subsequently, the machine-learning algorithms, including Liblinear, REPTree, XGBoost, MultilayerPerceptron, RandomTree, and RBFNetwork were applied to obtain the optimal model for video emotion recognition based on a multi-modal dataset. We discovered that using the data fusion of all-band EEG power spectrum density features and video audio-visual features can achieve the best recognition results. The video emotion classification accuracy achieves 96.79% for valence (Positive/Negative) and 97.79% for arousal (High/Low). The study shows that this method can be a potential method of video emotion indexing for video information retrieval.

INDEX TERMS Affective computing, video, EEG, multimodal, signal processing.

I. INTRODUCTION

External multimedia stimulation, as a carrier of human emotion, exhibits rich affective experience. Built from physiological arousal, mood swing is an abstract and implicit symbolic psychological action and it provokes the multimodal psychophysiological reaction. Given the effectiveness and instantaneity of video influence on human and various psychophysiological cues affected by emotion fluctuations, many researchers have devoted themselves in studying the relations among EEG signal, multimedia features and human emotion, and explored methods to achieve computer affective intelligence [1].

However, researchers cannot reach an agreement on how videos influence human emotion and the mechanism of video emotion recognition. Moreover, the types of videos and human physiological features that are crucial for emotion analysis and how to use machine-learning methods to realize artificial affective intelligence have not been concluded.

This study was conducted to obtain the best combination of features from the EEG-video fusion features and the optimal emotion recognition model by the comparison of different machine-learning algorithms. The experiment was designed to investigate the emotional relation between videos and human perception. We extracted audio-visual features from videos and EEG features from participants to form a video emotion database. Liblinear, REPTree, XGBoost, MultilayerPerceptron, RandomTree and RBFNetwork were compared to obtain the optimal model. The research roadmap is shown in Figure 1. Our finding offers a promising method for video emotion recognition with multimodal data fusion.

The main contributions of this paper can be concluded in two aspects:

(1) EEG features and multimedia features of the video were extracted to build the fusion dataset;

The associate editor coordinating the review of this manuscript and approving it for publication was Bora Onat.

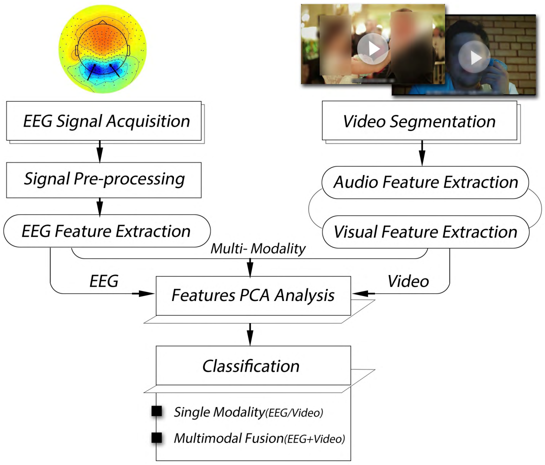


FIGURE 1. Research roadmap of the video emotion classification procedure.

(2) Liblinear, REPTree, XGBoost, MultilayerPerceptron, RandomTree, RBFNetwork algorithms were compared to build the optimal emotion recognition model. The video emotion classification accuracy achieved 96.79% for valence (Positive/Negative) and 97.79% for arousal (High/Low).

This paper is organized as follows: section 2 outlines the related studies in emotion recognition based on EEG features and multimedia features; section 3 describes the EEG measurement and video emotion annotation experiments; section 4 describes the affective EEG-video feature fusion dataset construction, including the feature extraction of all-band EEG signals and audio visual features in videos; section 5 presents a brief introduction of applied emotion recognition algorithms; section 6 provides the performance of different emotion recognition models; section 7 presents the discussion of the results; section 8 presents the conclusion and prospects of this study.

II. RELATED WORKS

Low level video information is important to retrieve emotional videos from massive multimedia data. In addition, the physiological features of the users, such as EEG, can also be used to study people's emotion response. Currently, some researchers have provided many useful methods to extract and analyze emotional features. The related works can be categorized into three classes based on applied feature modality: video features, EEG features and multimodal-EEG features fusion. Here we summarize the related studies of emotion analysis, including video low-level features types, induced user physiological feature categories, and the corresponding emotion classification accuracy.

A. EMOTION RECOGNITION BASED ON VIDEO FEATURES

The video emotion investigation study can be either affective video content analysis or video agent emotion recognition [2]. Video agent emotion recognition is aiming at estimating the emotion expressed by the agent or actors in videos. The database for these studies sometimes is formed by acted emotions of expressions and gestures [3]–[5]. The goal of video agent emotion recognition is to explore the method for human-computer affective interaction.

Affective video content analysis focuses on video emotion stimulus function, which is normally used as a stimuli material in the experiment. Thus low-level video features are extracted to map onto the emotion space to create affective representations [5]–[18]. The low level features of video primarily refer to the audio features and visual features extracted from videos, such as color, shape, sound, and texture. Many researchers have studied the mapping relationship from low level video features to a user's emotion space.

For both kinds of video emotion studies, a large number of early works with various machine learning algorithms have been performed, which can provide useful methods in video emotion analysis (listed in Table 1).

Most of the other studies that extracted audio and visual cues achieved a classification rate of 70%∼90% [8]–[10], [12], [13], [15], [16], [18], [19]. Specifically, Zhang et al. extracted audio and visual features to build a video emotion model with affinity propagation and achieved an accuracy of 92.9% for arousal and 89.2% for valence [10]. Y. Cui et al. applied SVR and MLR on a database of 552 video clips; the mean absolute error rate is 0.340 for arousal and 0.277 for valence [12]. Kang et al. used visual features to classify videos of fear, sadness and joy with an accuracy of 79%. Wang et al. performed an experiment with 2040 video clips and classified them in seven emotion classes with an accuracy of 74.69% [7]. For continuous emotion analysis, multimodal features were applied in the modeling, including low-level audio-visual features, facial features and textual features, which is important in video emotion recognition [11], [14], [15], [17].

The most widely used audio-visual features include visual features of color, lighting, motion, shot cut rate; audio features of MFCC, zero crossing rate, rhythm, sound energy, tempo and pitch etc. For instance, Kang [20] detected features from a video by extracting the color, motion and shot cut rate of the video to analyze their emotional states with AdaBoost algorithm. Hanjalic and Xu [21] believed that the motion, sound track and shot length of videos could reflect a lot of emotional information. They built a model to achieve the emotional annotation of the video. Zhang et al. [22] used arousal-pleasure as the emotional dimension to express music video emotion. They found that features of motion intensity, short switch rate, zero crossing rate, tempo and beat strength were closely related to arousal, and features of brightness, saturation, color energy, rhythm regularity and pitch were suitable for pleasure recognition. Wang et al. [23] used a semi-supervised learning mechanism to identify emotional information in the video. They extracted audio visual features from videos, using bivariate correlation analysis to select sensitive features, and subsequently applied a low-density separation algorithm to construct a classifier. They extracted

TABLE 1. Video affective analysis studies based on video features.

References	Features	Classifiers	Descriptors	Video clips	Annotators	Results
Kang,2003[6]	Visual	HMMs	Fear, Sadness, Joy	from six 30min videos	10	Accuracy: 79%
Wang and Cheong,2006[7]	Audio Visual	Two SVMs	7 Emotional Classes	2040	3	Accuracy: 74:69%
Sun and Yu, 2007[8]	Audio Visual	4 HMMs	(Anger, Fear, Sadness, Joy)	10	30	Precision:68% Recall: 79%
Irie et al.,2008[9]	Audio Visual	Latent Topic Driving Model	9 Emotion Categories	206	16	Agreement: 85.5%
S. Zhang et al.,2008[10]	Audio Visual	Affinity Propagation	Arousal, Valence	156	11 (1F,10M)	Accuracy: Precision: Arousal:0.93 Valence:0.89
Soleymani et al.,2009[11]	Textual	Bayesian	3 Emotional Classes	21 Full-length popular movies	1	Accuracy: 64%, F1:63%
Y.Cui et al., 2010[12]	Audio Visual	SVR and MLR	Arousal, Valence	552	11	Mean absolute error: Arousal: 0.340, Valence: 0.277
Xu et al.,2008[13]	Audio Visual	5 HMMs	5 Classes	Videos from 24 movies	N/A	Accuracy: 80.7%
Malandrakis et al.,2011[14]	Features at frame level	Two HMMs	Time series of 7 Categories Interpolated into Continuous	Video clips from 12 movies	7	Correlation for Arousal: 0.54 Valence: 0.23
Nicolaou et al.,2011[3]	Facial expression Shoulder gesture Audio	LSTM-RNN	Continuous recognition of emotions	SAL database	4	Correlation for Arousal:0.642 Valence:0.796
Kahou et al. ,2013[15]	AudioVisual	Combination of Multiple Deep Neural Networks	7 Emotional Classes	Short video clips extracted from movies	2	Accuracy: 41.03%
Acar et al., 2014[16]	Audio Visual	Two CNNs and one SVM	Arousal, Valence	DEAP dataset	32	Accuracy: 52.63%
Mandar Gogate et al.,2017[17]	Audio(6373 features) Text Visuals	CNN	Angry Happy Sad Neutral	IEMOCAP dataset	N/A	Angry90.1% Happy80.2% Sad76.4% Neutral78.3%
Baohan Xu,2018[18]	Image Transfer Encoding Audio	CNN	Probability of Correctly Guessing the Emotion	60 (2-second)	10	Accuracy of video summary : 4.1(The full score is 5.)
R. D. Fonnegra, 2018 [5]	Facial Expression	MLP	Facial emotion recognition	eNTERFACE'05 database	N/A	Accuracy: 81.84%

(HMM: Hidden Markov Model, CNN: Convolutional Neural Network, LSTM: Long Short-Term Memory, SVM: Support Vector Machine, SVR: Support Vector Regression), MLP: Multilayer Perceptron

69-dimensional audio features and used lighting and color energy to represent visual emotional features. Matsuda et al. [24] developed the EmoTour mobile application based on the data fusion of audiovisual data and human behavioral cues. They applied the RNN-LSTM algorithm on several existing emotion dataset to build the model and achieved an unweighted average recall of 0.484 by decision-level fusion. Moreover, emotion recognition is also used for video similarity analysis. J. Niu et al. presented a novel affect-based model of similarity measure of Internet videos based on emotion content. The emotion video dataset they applied was Affivir, which was established in their previous work [25], that included motion, shot-change rate, sound energy and audio pitch average features. The experimental results demonstrated the superiority of the proposed model [26]. More related studies and state-of- the-art methods were concluded in the surveys [2], [27].

The current studies of video emotions generally focus on the extraction of emotional video features and the modeling of emotional models. However, a semantic gap still exists between the low-level features of video and the individual emotions of users. Simultaneously, the emotional computation of physiological signals brings new opportunity to video emotion analysis. In this study, we will combine the audio visual features of video with the EEG features of participants to train the classifier.

B. EMOTION RECOGNITION BASED ON EEG fEATURES

The participants have instant and spontaneous emotional physiological responses during the time of videos watching. Picard analyzed four types of physiological signals, namely EMG, SC, RSP and BVP, and they extracted physiological features for identifying emotional states, proving the feasibility of using physiological signals for emotion recognition. Detecting human EEG information is another effective method according to the existing studies. However, the use of EEG information to model video emotion is still challenging owing to the limitation of datasets and complexity of EEG signal processing. Some of the studies on video emotion recognition based on EEG features are listed in Table 2.

In the studies of video emotion recognition based on EEG data, the power spectrum density (PSD), wavelet transform (WT) features, and fast Fourier transform (FFT) features are the widely used feature sets. Specifically, the PSD of various bands of EEG has been applied in numerous related works [28]–[34], [37], [38]. The experiment of EEG features collection typically requires a relatively longer time and more complex procedure comparing to the emotion annotation experiment without physiology signals detection. Consequently, the existing EEG dataset is insufficient owing to limited annotators.

Lin et al. [28] used support vector machine to classify EEG signals into four emotional states: joy, anger, sadness, and pleasure, subsequently, he discovered that the 30 independent characteristics of EEG were closely related to emotion. Nie [39] analyzed the frequency domain characteristics of EEG signals, and found that the signals of the delta (0-4Hz), theta (4-8Hz), alpha (8-13Hz), beta (13-30Hz) and gamma (30-45Hz) contribute significantly to emotion recognition, while the contribution of Theta and Alpha bands are low. Lin [40] performed a study using EEG in music video emotion recognition based on the theory of nonlinear dynamics, in which the nonlinear characteristics of the Shannon entropy, correlation dimension and C0 complexity extracted from EEG signals were utilized. The experimental results confirmed the feasibility of using the nonlinear dynamic analysis method to identify emotions. Furthermore, they found that the nonlinear dynamic EEG characteristics could improve the results significantly. A. Singhal et al. analyzed user emotion using EEG signals for video summarization through the crowd sourcing method. They achieved an average accuracy of 83.93% in the classification of happy, sad and neutral based on the random forest algorithm [41]. Barjinder Kaur et al. studied the emotion classification of calm, anger and happiness by EEG signals induced by videos and obtained an average accuracy of 60% using the support vector machine with a radial basis function [42]. Moon Inder Singh et al. developed a real time emotion classification method based on evoked EEG. The method achieved an accuracy of 55% on multiple subject trials [43]. Moreover, Jinyoung Moon et al. used interval EEG features of different bands combinations to detect user attention in video watching. The EEG features of the left hemisphere achieved the best F1 score of 39.60% with an average accuracy of 48.70%. They proposed that this method can be applied in intelligent video management application [44]. In recent studies, a key research problem is how to extract the feature vectors that are highly relevant with different emotional states from EEG signals. C. Shahnaz et al. presented a new emotion recognition method based on EEG wavelet transform features which is performed on the selected intrinsic mode functions [35]. Y. Li et al. provided the experiment results on different EEG features and proposed GRSLR model to achieve the best emotion recognition result [36]. Zhang et al. proposed a combined feature extraction method for EEG signals for emotion recognition [45]. Deep neural network is promising in selecting the crucial EEG frequency bands and channels to improve the emotion recognition accuracy [46]. More related studies and EEG feature extraction methods are concluded in the reviews [47]–[49].

All the previous studies indicate that measuring EEG signals is a possible method for emotion recognition. A key problem is how to extract the feature vectors that are highly relevant with different emotional states from EEG signals. The traditional quantitative analysis of EEG signals is based on the linear analysis method, which primarily analyzes its characteristics in the time domain and frequency domain. In recent years, some scholars studied the processing and analysis of EEG signals based on nonlinear dynamics theory. This study intends to use the frequency domain analysis method to extract the power spectrum density features of different frequency bands in EEG for emotion recognition.

C. EMOTION RECOGNITION BASED ON EEG-MULTIMODAL DATA FUSION

Apart from single-modal data, using multimodal data is also possible for video emotion recognition from both subjective and objective views. The combination of physiological

TABLE 2. Video affective analysis studies based on EEG data.

References	Features	Classifiers	Descriptors	Video clips	Annotators	Results
Y. Lin et al., 2010[28]	STFT, PSD	SVM,MLP	Joy, Anger, Sadness, Pleasure	16	26	Average Accuracy: 82.29% ±3.06%
D. Nie et al., 2011 [29]	FFT	SVM	Positive, Negative	12	6	Accuracy: 87.53%
X. Wang et al., 2013 [30]	PSD,WT, Nonlinear Dynamical feature	SVM	Valence, Arousal, Dominance,	12	6	Average Accuracy: 91.77%
Z. Lan et al., 2014 [31]	Fractal Dimension Feature, PSD, STATISTICS, HOC, ICC	SVM	Positive (Pleasant, Happy), Negative (Frightened, Angry)	4	4	Accuracy(different feature combinations): From 51.36%~91.67%
J. Atkinson et al.,2016 [32]	Maximum Relevance Minimum Redundancy Method Statistical	SVM RBF		DEAP database	32	2 Classes: Arousal 73.06%, Valence73.14% 3 Classes: Arousal60.7%, Valence 62.33%
A. M. Bhatti, et al.,2016[33]	Statistical, PSD, FFT, WT	MLP-BP	Happy, Sad, Love, Anger	Music	30	Happy94.87%, Love 65.38%, Sad78.13%, Anger 74.07%
E. Kroupi et al., 2016[34]	PSD, DFT, WD	LDA	Pleasantness	Odors	25	Pleasant57%, Unpleasant100%, Neutral 42%
C. Shahnaz et al., 2016[35]	WT	SVM	Valence, Arousal, Dominance, Liking	DEAP database	32	Valence74.94%, Arousal76.68%, Dominance76.67%, Liking81.9356%
Y. Li et al., 2018[36]	STATISTICS, PSD, ENTROPY, HOC, HJORTH	GRSLR	Happy, Neural，Sad	SEED database	14	Happy 84.34%, Neutral 80.96% , Sad 76.45%

(HMM: Hidden Markov Model, CNN: Convolutional Neural Network, LSTM: Long Short-Term Memory, SVM: Support Vector Machine, SVR: Support Vector Regression, MLP: MultilayerPeceptron; RBF: Radial Basis Function, BP: Back Propagation, GRSLR :Graph Regularized Sparse Linear Regression, LDA: Latent Dirichlet Allocation, PSD: Power Spectrum Density, SFT: Short-time Fourier Transform, WT: Wavelet Transform, FFT: Fast Fourier Transform, DFT: Discrete Fourier Transform, WD: Wasserstein Distance),HOC: Higher-Order Crossings, ICC: Intra-class Correlation Coefficient

signals and video contents will increase the complexity of feature analysis and modeling work. However it can consider the individual emotion perception characters to build a personalized emotion model.

The frameworks of emotion analysis based on EEG-multimodal data fusion have similarity among existing studies. Affective images, music [50], videos [51]–[57] have been selected as the stimuli to induce emotions. Users' emotion states are measured by physiological and EEG sensors and the emotion annotations are rated by users in the experiment accordingly. Finally different classifiers are applied in different combination of multimodal features dataset. The emotion prediction outcome can be influenced by many factors, including the selection of feature extraction method, experiment procedure design, participants' perception and algorithm modeling. Besides, most of the existing studies demonstrated better model performance in predicting valence by multimodal physiological features [53], [55]. And Koelstra [54] and Wiem et al. [58] presented that using EEG features has better effectiveness in arousal classification.

According to the recent literatures, emotion recognition based on EEG and video features has achieved a robust modeling performance [55], [56]. Emotion recognition based on

TABLE 3. Video affective analysis studies based on EEG-multimodal fusion data.

References	Features	Classifiers	Descriptors	Video clips	Annotat ors	Results
Takahashi, 2003[52]	EEG , Pulse, Skin Conductance	SVM	Joy, Sadness, Disgust, Fear, Relax	N/A	12	Accuracy 41.1%
Sander Koelstra, 2010[53]	EEG:Power Spectral Density (PSD) and Common Spatial Patterns (CSP).) GSR,EOG,HR,EMG, Blood Volume Pressure, Respiration, Skin Temperature	SVM	Arousal, Valence Like	20	6	EEG: Arousal 55.7% ; Valence 58.8% ; Physiological signals: Arousal 58.9% ; Valence 54.2%
Sander Koelstra, 2012[54]	GSR, Blood volume pressure, Respiration, Skin temperature, EMG, EOG, EEG Color variance, lighting key MFCC, Energy Formants, Pitch, Zero crossing rate, Silence ratio	Weighting Scheme	Arousal, Valence, Dominance	40	32	F1: Arousal 0.618 Valence 0.605 Liking 0.634
Mohammad Soleymani, 2012[55]	ECG, GSR, Respiration Amplitude, Skin Temperature, Eye Gaze, Facial Expressions,EEG,Audio	SVM RBF	Arousal, Valence	Experiment 1:20/ Experiment 2:14	27	Arousal 67.7% Valence76.1%
Mohammad, 2012[56]	EEG,pupil	SVM	Arousal, Valence	20	24	Arousal 76.4% Valence 68.5%
Mojtaba, 2015[51]	MEG,EOG,EMG,ECG,EEG,NIR Audio,video	NaiveBayes	Arousal, Valence, Dominance	20 Music/18 Movie clips	30	Arousal58.3% Valence50.0%, Dominance52.8%
Mohammad Soleymani, 2016[57]	Power Spectral Density Facial Expressions	MLR SVR LSTM-RNN	Arousal, Valence	239	28	RMSE: Arousal 0.143 Valence 0.185

Jia-Lien Hsu, 2018[50]	Music(Spectrum, Spectral Centroid, MFCCs,Ttonal, and Harmonic Timbre)
	EEG(EEG rhythms)
		ANN	Arousal,
			Valence
				IADS	23	Personalized model for music emotion recognition :RMSE:e0.176

L. Granados, 2018[59]	EEG, GSR	DCNN	Arousal, Valence	20	57	Arousal 71% Valence 75%
J. Cao et al., 2018[60]	EEG Audio, Visual	ELM	Violence, Neutral, Eroticism	90	6	Accuracy: 76.67%
T. Song et al., 2019[61]	EEG, ECG,GSR,RSP	A-LSTM	Joy, Funny, Anger, Fear, Disgust, Sad, Neutrality	170	162	Presented 3 kinds of protocols for emotion recognition

(HMM: Hidden Markov Model, CNN: Convolutional Neural Network, LSTM: Long Short-Term Memory, SVM: Support Vector Machine, SVR: Support Vector Regression, MLP: MultilayerPeceptron; RBF: Radial Basis Function, BP: Back Propagation, LDA: Latent Dirichlet Allocation，RNN: Recurrent Neural Networks, ANN: Artificial Neural Network ,DCNN:Deep Convolutional Neural Network, A-LSTM: attention-long short-term memory ,ELM: Extreme Learning Machine)

multimodal feature fusion is a promising method. The related studies are listed in Table 3.

Sander Koelstra et al. combined the physiological data (GSR, blood volume pressure, RSP, skin temperature, EMG, EOG and EEG) with audio-visual features to form a fusion emotion dataset, and the experiment result indicates that fusion dataset has the best performance in emotion recognition [52], [53]. Mohammad Soleymani et al. performed a series of studies on video emotion recognition based on multimodal physiological features [51], [56], [57], [62]. They were devoted to the research of multimodal emotion recognition and sought for the optimal model performance with the best feature sets. In 2012, they applied physiological signals in video emotion analysis, including skin electricity, blood pressure, respiration, EMG, eye movement, and temperature from audiences. They constructed a linear mapping from the physiological feature space to the arousal-pleasure space, and experimentally demonstrated the feasibility of using user physiological signals to implicitly annotate emotional video semantics. In 2016, they selected EEG PSD and facial expression features to build a video emotion model and achieved a rmse of 0.143 for arousal and 0.185 for valence.

In the recent study performed by Jia et al. in 2018, music features and EEG features were combined to fuse the dataset and build the personalized model for music emotion recognition, which achieved an average mse of 0.176 [50]. L. Granados et al. applied the deep learning approach on the EEG and galvanic skin response dataset and achieved a classification accuracy of 71% for arousal and 75% for valence [59]. L. Duan et al. presented a video-EEG features fusion approach using kernel-based Extreme Learning Machine for video emotion recognition and achieved an accuracy of 76.67% [60]. T. Song et al. built a multi-modal physiological emotion database for discrete emotion recognition (MPED). They investigated the emotion recognition of multimodal physiological signals with three types of emotion classification protocols and compared different feature set performance with different algorithms [61].

The various methods also indicate the possibility of using multimodal data fusion in affective video computing, and the model performance is becoming increasingly competitive in recent studies. Thus, multimodal data fusion is the promising method in this research field. As shown, in the present studies, the recognition model performance differed (ranging from 40% to 80%), owing to the dataset limitation, stimuli material and annotation variation. Accordingly, the universal problem of the existing studies is that there is no standard specification for standardizing databases, stimulating materials, methods of extracting features, definition of features weight, and selection of model algorithms. The sample size of the dataset is relatively small, it is difficult to use algorithms based on the large database such as DeepCNN.

Herein, we review the research on emotion feature extraction and the analysis from video low level features and user physiological features, and summarize the machine-learning classification methods typically used in emotion computing. From the literature review, we found that the extraction and classification of emotional features have not yet formed a systematic theoretical system. Video emotion recognition technology still exhibits great potential for further study. In this study, a video emotion classifier based on EEG will be constructed. Specifically, we compared several machine-learning algorithms in emotion classification to obtain the optimal model.

III. EXPERIMENT

To our knowledge, the quantity of studies that using EEG and multimedia feature fusion to recognize video emotion is limited; therefore, the EEG emotion video dataset for affective computing is inadequate currently. Therefore, an original EEG-video emotion dataset is demanded to obtain the EEG signals produced by the participants induced by the video stimulation.

A. EEG SIGNAL COLLECTION EQUIPMENT

We applied the ActiveTwo system of Biosemi to measure the EEG signal of the participants. BioSemi provides a range of state-of-the-art equipment for physiological signals measurements. BioSemi is currently one of the top EEG hardware companies in the world [63]. The system contains 66 channels, in which 64 channels are used to collect EEG signals and the other two channels are used to collect eye movement signals for denoising and calibration [64].

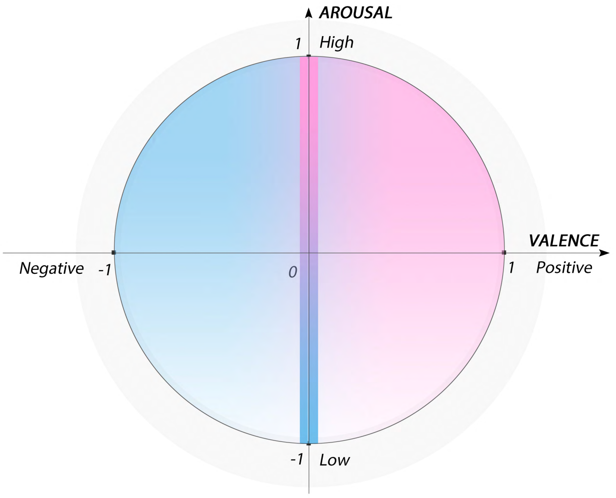


FIGURE 2. Valence-arousal emotion model [1].

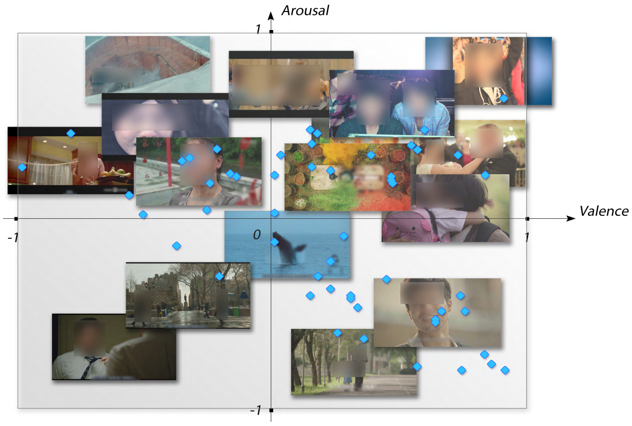


FIGURE 3. Video emotion distribution mapping.

B. AFFECTIVE STIMULI MATERIALS

A total of 71 video clips were employed to form the emotion stimuli dataset in the experiment. These videos were selected from the internet videos with various emotion categories distributed in the valence-arousal emotion model space (see Figure 2).

The emotion class expressed in each video clip is shown in Figure 3. Each point represents a video clip emotion score. It is ensured that the video materials cover all four quadrants in the valence-arousal emotion space that render the dataset balanced in each dimension.

The duration of each video clip is 50 s, which was typically selected from the principal part of the video with a primary emotion to prevent the emotion cognition confusion. They

TABLE 4. Source of the affective stimuli videos.

Emotion	Videos		The sources of stimuli video clips

A	H	43	Life of Pi;
			The most wonderful moment: Wedding;
			Amazing, my country;
			Extreme sports video

	L	10	Global Warming;
			Thailand Touching Advertising Videos Collection;
			Hachi: A Dog's Tale;
			Spring Festival travel rush

V	P	30	Amazing, my country;
			The most wonderful moment: Wedding;
			Animal World;
			Extreme sports video

	N	23	Life of Pi;
			Global Warming;
			Thailand Touching Advertising Videos Collection;
			Hachi: A Dog's Tale;
			Spring Festival travel rush;

(A:Arousal,V:Valence,H:High,L:Low,P:Positive,N:Negative)

TABLE 5. Experiment participants and data collection of each group.

Group	Participants	Video clips	EEG instances
1	10	17	170
2	10	19	190
3	10	16	160
4	9	19	171
Total	**39**	**71**	**691**

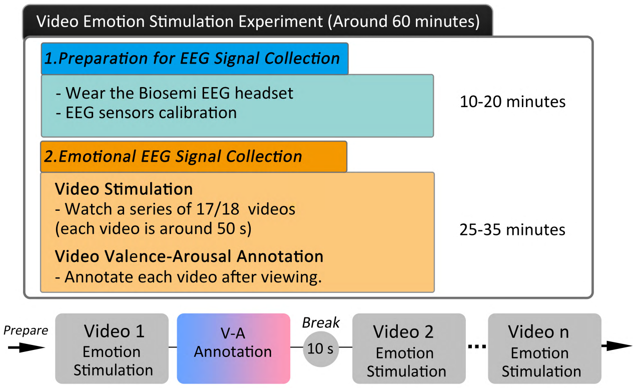

Video Emotion Stimulation Experiment (Around 60 minutes)
1.Preparation for EEG Signal Collection
- Wear the Biosemi EEG headset
- EEG sensors calibration
10-20 minutes
2.Emotional EEG Signal Collection
Video Stimulation
- Watch a series of 17/18 videos
(each video is around 50 s)
Video Valence-Arousal Annotation
- Annotate each video after viewing.
25-35 minutes
Prepare
Video 1 Emotion Stimulation
V-A Annotation
Break
10 s
Video 2 Emotion Stimulation
...
Video n Emotion Stimulation

FIGURE 4. Emotion stimulation experiment process.

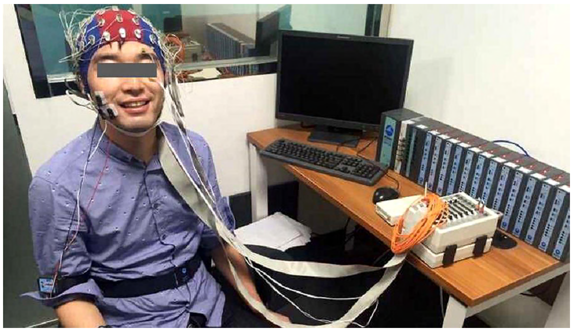


FIGURE 5. EEG measurement experiment induced by affective videos.

were cut manually from different Internet videos by students majoring in movie production. We selected videos that contain music and scene without dialogue to prevent semantic misleading on the participants. The sources of the affective stimuli videos are listed in Table 4.

People have subjective perception and perspectives for the same video, which could cause diversity in emotion scoring. Videos without cognition consistency were abandoned in the experiment. Finally, there were 30 positive videos (valence>0) and 23 negative videos (valence<= 0) distributed in the valence dimension, 43 high-arousal data (arousal>0) and 10 low-arousal data (arousal<= 0) distributed in the arousal dimension.

C. PARTICIPANTS

A total of 39 participants were enrolled to score 71 videos; and then subsequently 53 videos with emotion labeling consistency were selected for the next session of the experiment. A total of 691 EEG instances were recorded in the experiment and 499 of them were valid after evaluation. The arrangement of the participants' group and the number of EEG instances collected from each group are listed in Table 5.

D. EXPERIMENTAL DESIGN AND PROCEDURE

The participants were arranged into four groups. The experiment was designed as several steps. The specific arrangement is presented in Figure 4.

The detailed process of video emotion annotation experiment is shown as below:

- Thirty-nine volunteers (aged 18 to 35 years old) were invited to participate in the experiment.
- Each participant was asked to read the instructions carefully for the experiment workflow and procedure and sign the agreement of experiment. Then he/she entered the laboratory room and wore the sensors of EEG Biosemi system. The laboratory researcher adjusted the sensors' working status to ensure the signal record quality.
- The experiment was conducted in a separate room with constant lighting and temperature setting. The participant was seated in front of a 21-inch screen to watch the video during the experiment, see Figure 5.
- The participants wore the sensors of Biosemi equipment with 66 channels, in which 64 channels were used for EEG measurement and two channels were eye tracking measurement for EEG signals calibration.

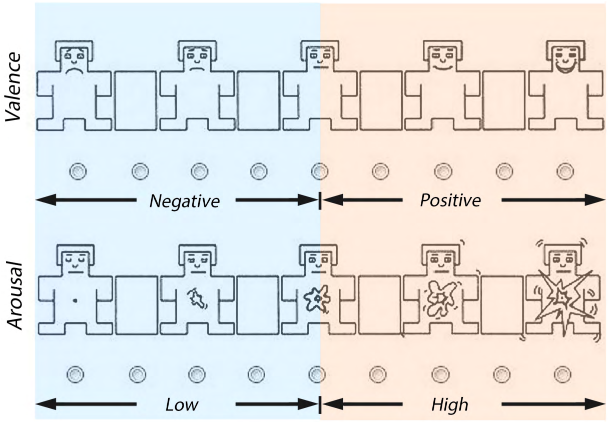


FIGURE 6. Valence-arousal emotion labeling questionnaire [49].

- Video watching session: Each participant watched one group of 17 or 18 video clips (see Table 1), with a 10 s break between two videos.
- Emotion annotation session: The participant annotated in valence-arousal dimension (scoring on a scale of 1 to 9) on the emotion labeling questionnaire [65] (see Figure 6) after watching a video. The whole experiment lasted for about 60 minutes.
- The participant got 100 RMB for his/her efforts as a reward after the experiment.

E. ANNOTATION ANALYSIS

The video emotion annotations are analyzed in the process below:

- If the emotion scored by any participant has much disparity with others, the whole physiological data instance and the corresponding scores will be abandoned. For example, the videos with 6 annotations of high positive high arousal and 4 annotations of high negative low arousal in the experiment will be confirmed that it might generate various interpretations. This situation is common in moving videos, which describes positive contents but convey sad emotion;
- The video with controversial emotion scores will be considered as confusing emotion stimuli material and eliminated from the dataset;
- The average value of the emotion score (valence-arousal) is set to be the emotion label of the video.

IV. SIGNAL PROCESSING AND FEATURE EXTRACTION

In this section, we propose the detailed feature extraction methods of EEG signal features and video features. The procedure of the multimodal features extraction is shown in Figure 7.

A. EEG FEATURES EXTRACTION

EEG Data Pre-processing:

EEG data are the electrical signals of brain activities, which is highly sensitive to interference of eye blinks and facial muscle activities. Thus, EEG signal pre-processing is crucial for noise elimination. In this study, we used the EEGLAB [66] Analyzer software to preprocess the EEG signals:

- First, the reference electrode was set;
- Second, the signals of the eye track channels were used to remove the eye movement influence on signals in the EEG channels;
- Third, we conducted EEG signal filtering, segmentation, baseline calibration, and removed artifacts;
- Finally, the EEG signal data for feature extraction was obtained by data signal averaging superposing.

EEG Features Extraction:

In this study, the traditional frequency-domain analysis method was utilized to extract the PSD features of 64 channels of EEG signals in five different frequency bands. The EEGLAB toolbox of MATLAB was applied. The five different frequency bands are delta $(0-4Hz)$, theta $(4-8Hz)$, alpha $(8-13Hz)$, beta $(13-30Hz)$ and gamma $(30-45Hz)$ [39].

The PSD features extraction method [38], [67] is shown as following:

Denote the ith signal x windowed, then zero-padded frame from the signal x by

$$x_i(k) \triangleq w(k)x(k+i\theta), \quad k = 0, 1, \ldots, I-1, \quad i = 0, 1, \ldots, Q-1 \tag{1}$$

where, θ is set as the window hop size, and Q is the number of available frames. So that the periodogram of the ith block is defined as

$$P_{x_i,I(\omega_q)} = \frac{1}{I}|FFT_{K,q}(x_i)|^2 \triangleq \frac{1}{I}|\sum_{k=0}^{K-1} x_i(k)e^{-j2\pi kq/K}|^2 \tag{2}$$

Then the power spectral density is given by

$$\hat{S}_x^D(\omega_q) \triangleq \frac{1}{Q}\sum_{i=0}^{Q-1} P_{x_i,I(\omega_q)}. \tag{3}$$

Finally, 320 dimensions of features were extracted in this experiment.

B. VIDEO FEATURES EXTRACTION

In this study, auditory and visual features were extracted from videos to form the video features dataset. The specific feature description and extraction methods are introduced below.

Audio Features Extraction

Regarding audio feature extraction, we utilize the openSMILE toolkit [4] to extract the 88 baseline acoustic features with the extended Geneva minimalistic acoustic parameter set (eGeMAPS) [68] configuration file, including loudness, MFCCs, F0, shimmer, spectrum flux etc., see Table 6.

Visual Features Extraction

In terms of visual feature extraction, we applied MATLAB to extract two-dimensional audio emotional features, including the luminance coefficient and color energy [7], see Table 7.

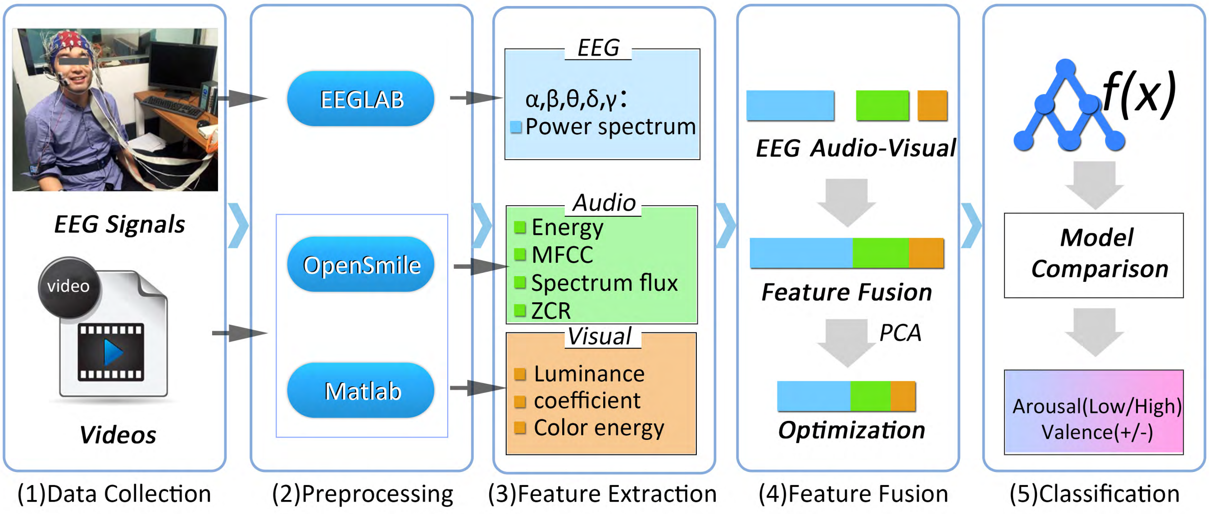


FIGURE 7. Procedure of multimodal features extraction.

To our knowledge, the brightness of the video is important in affecting the mood of the audience and the rendering atmosphere in video scenes. Herein, the luminance coefficient of the video is extracted as one of the low-level visual features, which is calculated as the value calculated by the luminance mean and variance of each frame.

$$\overline{v} \times \text{var} \tag{4}$$

In which, the average value of the luminance is $\overline{v} = \sum_{k=1}^{M} v_k$, the variance of the luminance is $\text{var} = \sum_{k=1}^{M} (v_k - \overline{v})^2$,$v_k$ represents the luminance of no. k pixel point, M represents the total number of pixel points of the frame image.

Arousal is closely related to luminance, while valence is closely related to color saturation [37]. Consequently, the color energy is selected as another visual feature, as it could reflect the saturation, luminance and region color. The color energy is the value related to color contrast and energy.

$$\sum_i \sum_j p(c_i)p(c_j)d(c_i, c_j) \times \sum_{k=1}^{M} E(h_k)s_k v_k \tag{5}$$

Here, h_k, s_k, $v_k$represent the color degree, saturation and luminance respectively of the no.k pixel point. $E(h_k)$ is the chroma energy. $\sum_i \sum_j p(c_i)p(c_j)d(c_i, c_j)$ represents the color contrast degree. c is the value of HSL color histogram bin. $d(c_i, c_j)$ represents the Euclidean distance between c_i and c_j in HSL space; M is the total number of the pixel points.

V. EMOTION RECOGNITION MODELING ALGORITHMS

In this study, we used Python to build the algorithm models with 10 folds cross-validation. In the experiment, five algorithms were implemented for the classification of valence (positive/negative) and arousal (low/high), including Liblinear, REPTree, XGBoost, MultilayerPerceptron, RandomTree and RBFNetwork. A brief introduction of these algorithms is presented below.

Liblinear

Liblinear is a linear classifier for large dataset of instances and multidimensional features. It supports classifiers of Linear Regression and Support Vector Machine, including L2-regularized classifiers, L2-loss linear SVM, L1-loss linear SVM, and logistic regression (LR), L1-regularized classifiers (after version 1.4), L2-loss linear SVM and logistic regression (LR), L2-regularized support vector regression (after version 1.9), and L2-loss linear SVR and L1-loss linear SVR. For a set of instance (m_i, n_i), $i = 1, \ldots, k$, $m_i \in D^n$, $n_i \in \{-1, 1\}$, this method solves the unconstrained optimization problem:

$$\min_{\varepsilon} \frac{1}{2}\varepsilon^T \varepsilon + \rho \sum_{i=1}^{k} \gamma(\varepsilon; m_i, n_i) \tag{6}$$

in which ρ >0 is defined as a penalty parameter and $\gamma(\varepsilon; m_i, n_i)$ is the loss function. The experiments prove that it is highly efficient in the classification of large data sets [69].

REPTree

REPTree is a fast decision tree learner. It uses information gain/variance to build a decision or regression tree, and prunes it using reduced-error pruning with backfitting. This method reduces the complexity of decision tree method. Given two distinct variables X and Y with values $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_n\}$, then the entropy and conditional entropy of Y is defined as:

$$E(Y) = -\sum_{i=1}^{k} U(Y = y_i) \log U(Y = y_i) \tag{7}$$

$$E(Y|X) = -\sum_{i=1}^{t} U(X = x_i) \log U(Y|X = x_i) \tag{8}$$

Then the information gain of $X$is given by:

$$\varphi(Y; X) = E(Y) - E(Y|X) \tag{9}$$

The pruning can be done with pre-pruning and post-pruning. In the pre-pruning process, if the information gain due to

TABLE 6. Audio features extracted from the videos.

Feature Category	Detailed features	Dimensions
Loudness	amean stddevNorm percentile20.0 percentile50.0 percentile80.0 pctlrange0-2 meanRisingSlope stddevRisingSlope meanFallingSlope stddevFallingSlope loudnessPeaksPerSec	11
F0semitone From27.5Hz	amean stddevNorm percentile20.0 percentile50.0 percentile80.0 pctlrange0-2 meanRisingSlope stddevRisingSlope meanFallingSlope stddevFallingSlope	10
MFCC	mfcc1_sma3_amean mfcc1_sma3_stddevNorm mfcc2_sma3_amean mfcc2_sma3_stddevNorm mfcc3_sma3_amean mfcc3_sma3_stddevNorm mfcc4_sma3_amean mfcc4_sma3_stddevNorm mfcc1V_sma3nz_amean mfcc1V_sma3nz_stddevNorm mfcc2V_sma3nz_amean mfcc2V_sma3nz_stddevNorm mfcc3V_sma3nz_amean mfcc3V_sma3nz_stddevNorm mfcc4V_sma3nz_amean mfcc4V_sma3nz_stddevNorm	16
jitterLocal	jitterLocal_sma3nz_amean jitterLocal_sma3nz_stddevNorm	2
shimmerLocal	shimmerLocaldB_sma3nz_amean shimmerLocaldB_sma3nz_stddevNorm	2
Spectrum flux	spectralFlux_sma3nz_amean spectralFlux_sma3nz_stddevNorm spectralFluxV_sma3nz_amean spectralFluxV_sma3nz_stddevNorm spectralFluxUV_sma3nz_amean	5
HNRdBACF	HNRdBACF_sma3nz_amean HNRdBACF_sma3nz_stddevNorm	2
logRelF0-H1-H2	logRelF0-H1-H2_sma3nz_amean logRelF0-H1-H2_sma3nz_stddevNorm logRelF0-H1-A3_sma3nz_amean logRelF0-H1-A3_sma3nz_stddevNorm	4
	F1frequency_sma3nz_amean F1frequency_sma3nz_stddevNorm F1bandwidth_sma3nz_amean F1bandwidth_sma3nz_stddevNorm F1amplitudeLogRelF0_sma3nz_amean F1amplitudeLogRelF0_sma3nz_stddevNorm	

TABLE 6. Audio features extracted from the videos.

Frequency	F2frequency_sma3nz_amean F2frequency_sma3nz_stddevNorm F2bandwidth_sma3nz_amean F2bandwidth_sma3nz_stddevNorm F2amplitudeLogRelF0_sma3nz_amean F2amplitudeLogRelF0_sma3nz_stddevNorm F3frequency_sma3nz_amean F3frequency_sma3nz_stddevNorm F3bandwidth_sma3nz_amean F3bandwidth_sma3nz_stddevNorm F3amplitudeLogRelF0_sma3nz_amean F3amplitudeLogRelF0_sma3nz_stddevNorm	18
alphaRatio	alphaRatioV_sma3nz_amean alphaRatioV_sma3nz_stddevNorm alphaRatioUV_sma3nz_amean	3
hammarbergIndex	hammarbergIndexV_sma3nz_amean hammarbergIndexV_sma3nz_stddevNorm hammarbergIndexUV_sma3nz_amean	3
slopeV0-500	slopeV0-500_sma3nz_amean slopeV0-500_sma3nz_stddevNorm slopeV500-1500_sma3nz_amean slopeV500-1500_sma3nz_stddevNorm slopeUV0-500_sma3nz_amean slopeUV500-1500_sma3nz_amean	6
VoicedSegment	VoicedSegmentsPerSec MeanVoicedSegmentLengthSec StddevVoicedSegmentLengthSec MeanUnvoicedSegmentLength StddevUnvoicedSegmentLength	5
equivalentSoundLevel	equivalentSoundLevel_dBp	1
Total		**88**

division is not recognized, the expansion of the tree will be terminated [70].

XGBoost

Extreme gradient boost (XGBoost) is an optimized distributed gradient boosting algorithm with high efficiency and flexibility. It is implemented under the gradient boosting framework. XGBoost can be used in solving supervised learning problems (classification and regression) rapidly and accurately. Given the training data a_i to predict the variable b_i, the method is defined as:

$$\hat{b}_i = \sum_{q=1}^{Q} f_q(a_i), \quad f_q \in U \tag{10}$$

where Q is the number of trees, f_q represents a function of the qth tree in the function space U, which is the set of all possible CARTs [71].

MultilayerPeceptron

Multilayer perceptron (MLP) is a class of supervised learning algorithm of feedforward neural network. It consists of multiple layers, including an input layer, a hidden layer and an output layer, that distinguish MLP from a linear perceptron. It can distinguish data that are not linearly separable for either classification or regression. Given an input signal x_i and its output y_i is computed as a function defined as:

$$y_i = f(\sum \beta_{ji} x_i) \tag{11}$$

where β_{ji} is the connection weight, f is the activation function. MLP is a popular machine-learning solution that is widely used in multimedia emotion recognition. [72].

TABLE 7. Visual features extracted from the videos.

Feature Category	Detailed features	Dimensions
Luminance coefficient	Lighting Key	1
Color energy	Color Energy	1

RandomTree

Random Tree constructs a tree that considers K randomly chosen attributes at each node, which is formed by a stochastic process. It does not perform pruning. In addition, every parameter on trees is then a random variable, for instance, the number of leaves or the diameter. RandomTree also has an option to allow for the estimation of class probabilities, or target mean in the case of regression, based on a hold-out set [73].

TABLE 8. Emotion classification accuracy based on EEG data.

Algorithm	Valence		Arousal		*Parameters setting*
	F1	Accuracy	F1	Accuracy	
Liblinear	0.487	50.70%	0.572	59.12%	L2-regularized L2-loss support vector (dual), cost 1, eps 0.001
REPTree	0.386	54.51%	0.490	60.72%	minNum 2, minVarianceProp 0.001, numFolds 3, seed =3(for Valence); 1(for Arousal)
XGBoost	0.546	50.00%	0.6562	56.00%	seed = 8,test_size = 0.3
MLP	0.485	48.70%	0.498	58.52%	learningRate 0.3, momentum 0.2, validationThreshold 20, seed =1(for Valence); 0(for Arousal)
RandomTree	**0.546**	**54.71%**	**0.500**	**62.32%**	minVarianceProp 0.001, numDecimalPlaces 2, maxDepth=0, minNum=2,seed=1(for Valence) maxDepth=1, minNum=1,seed=0(for Arousal)
RBFNetwork	0.504	50.50%	0.533	61.52%	minStdDev 0.1, numClusters 2, ridge 1.0E-8 clusteringSeed= 0, maxIts=1(for Valence) clusteringSeed= 1, maxIts=-1(for Arousal)

RBFNetwork

The radial basis function network (RBFNetwork) is an artificial neural network that implements a normalized Gaussian RBFNetwork. The network consists of three layers, including an input layer, a hidden layer (RBF units) and an output layer. The hidden layer is activated by a radial basis function, which can be described by $\beta_i : X_n \rightarrow X$. It uses the k-means clustering algorithm to learn either a logistic or linear regression. By this means, it reduces the number of dimensions of input data and transforms the data to a new space. Then an optimal estimation of the kernel parameter is conducted. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters [74].

VI. VIDEO EMOTION RECOGNITION MODELING

In this study, EEG features and video features were used to form the fusion database for modeling. Subsequently, Liblinear, REPTree, XGBoost, MultilayerPerceptron, RandomTree and RBFNetwork were used in the recognition modeling to evaluate the classification results on a single modal dataset and fusion dataset. The F-measure score and classification accuracy were chosen to compare the algorithms performance.

A. VIDEO EMOTION DATASET

The video emotion dataset employed in this experiment consists of 53 video clips that are cut manually from Internet videos with balance emotion labels. Each video clip lasts approximately 50 s with an audio sampling rate of 44100Hz. A total of 39 participants were invited to annotate the video clips in the valence-arousal emotion model. Furthermore, the participants' EEG data, video features and V-A emotion labeling scores were collected to form the video emotion database.

B. VIDEO EMOTION RECOGNITION BASED ON SINGLE MODALITY

Firstly, we used the feature set of the EEG signals and video feature separately to test the recognition results of single modality. In each dataset, the PCA algorithm was applied to reduce the feature dimensions. A comparison experiment was conducted with Liblinear, REPTree, XGBoost, Multilayer-Perceptron, RandomTree and RBFNetwork to obtain the best model for both single modality datasets. In this study, 10 folds cross validation was used to evaluate the performance of the classifiers. The model results and parameters setting are presented in Tables 8 and 9, respectively. It is shown that MLP achieves the lowest F-measure and the highest classification accuracy, and achieves the best result compared with other algorithms for both datasets.

C. EMOTION RECOGNITION BASED ON MULTIMODALITIES DATA FUSION

In this study, we combined the EEG and audio visual features of the video; subsequently, the PCA algorithm was applied to reduce the feature dimensions. A comparison experiment was conducted with Liblinear, REPTree, XGBoost, Multilayer-Perceptron, RandomTree and RBFNetwork to obtain the best model. The model results are presented in Table 10, and MLP with the lowest F-measure and highest classification accuracy achieves the best result.

It can be concluded from the results in Table 6 that most of the algorithms exhibit better classification effect on arousal than valence. This result is consistent with those in most related studies [28], [42], [43]. This might be because the positive and negative emotions evoked by videos are influenced by the subjective understanding of the participants; nevertheless, the participants exhibit the similar recognition on the arousal degree of the videos.

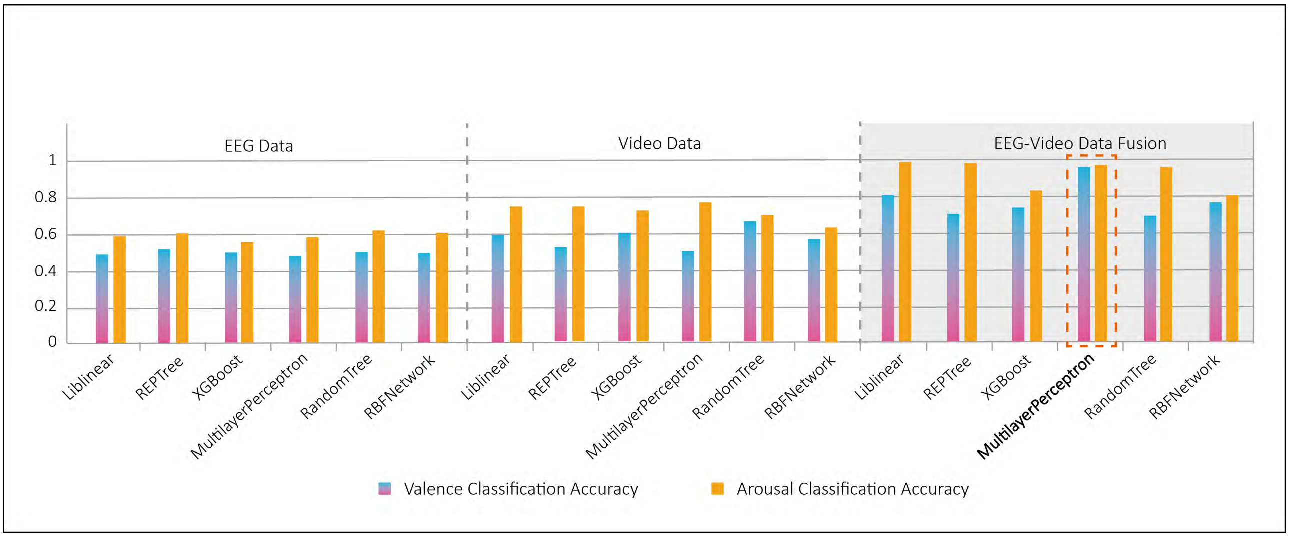


FIGURE 8. Emotion classification accuracy comparison based on EEG data/video data/EEG-video data fusion.

TABLE 9. Emotion classification accuracy based on video data.

Algorithm	Valence		Arousal		*Parameters setting*
	F1	Accuracy	F1	Accuracy	
Liblinear	0.600	60.38%	0.702	75.47%	L2-regularized logistic regression(primal), eps 0.001, cost =1(for Valence); 2(for Arousal)
REPTree	0.353	54.71%	0.430	75.47%	minNum 2, minVarianceProp 0.001, numFolds 3, seed=2(for Valence); 1(for Arousal)
XGBoost	0.667	60.00%	0.8462	73.33%	seed = 5,test_size = 0.3(for Valence); 0.33(for Arousal)
MLP	**0.529**	**52.83%**	**0.758**	**77.36%**	learningRate 0.3, momentum 0.2, validationThreshold 20, seed=1(for Valence); 0(for Arousal)
RandomTree	0.678	67.92%	0.640	73.58%	maxDepth 1, minNum 2, minVarianceProp 0.001, numDecimalPlaces 2, seed 1
RBFNetwork	0.579	58.49%	0.609	64.15%	maxIts -1, minStdDev 0.1, numClusters 2, ridge 1.0E-8, clusteringSeed=2(for Valence); 1(for Arousal)

The result of this experiment section shows that the fusion dataset exhibits better performance than the single modal features in terms of classification accuracy both in the arousal and valence dimensions. It is shown that when the EEG and video features are used separately to train the classifiers, they present the similar recognition result in the same emotional dimension. The comparison of classification based on different modalities is shown in Figure 8.

According to the model comparison results in Table 10, MultilayerPerceptron outperforms the other algorithms in average accuracy of arousal and valence classification. Liblinear and REPTree have better performance in arousal recognition, however they achieve relatively lower accuracy in valence classification. The classification accuracy of MLP for arousal achieves 97.79%, and the classification accuracy for valence achieves 96.79%, which is a balanced classification result.

In the model parameter setting process, we set the value of *seed* of MLP from 0 to 10 with step 1 to obtain the best performance for arousal and valence, and *LearningRate* is set from 0.1 to 1 with step 0.1 to find the best score setting. The parameter exploration is presented in Figures 9 and 10, that when $seed = 6$ for valence and 0 for arousal, while $LearningRate = 0.4$ for valence and 0.3 for arousal, MLP achieves the best result in the modeling work.

It is interesting to find that although the recognition rate of the EEG features and video features is not desirable, the data fusion can serve a complementary role, and thus significantly improves the performance of video emotion classification

TABLE 10. Emotion classification accuracy based on EEG-video data fusion.

Algorithm	Valence		Arousal		Parameters setting
	F1	Accuracy	F1	Accuracy	
Liblinear	0.806	80.56%	0.992	99.19%	cost 1, eps 0.001 support vector by Cranmmer and Singer(for Valence) L2-regularized L2-loss support vector (primal)(for Arousal)
REPTree	0.706	70.74%	0.986	98.59%	minNum 2, minVarianceProp 0.001, numFolds 3, seed=2(for Valence); 1(for Arousal)
XGBoost	0.782	74.00%	0.905	83.33%	seed = 4,test_size = 0.3
MLP	**0.968**	**96.79%**	**0.978**	**97.79%**	learningRate 0.4(for Valence);0.3(for Arousal), momentum 0.2, validationThreshold 20, seed=6(for Valence); 0(for Arousal)
RandomTree	0.698	69.74%	0.966	96.59%	maxDepth 0, minVarianceProp 0.001, numDecimalPlaces 2, seed 1, minNum=2(for Valence); 1(for Arousal)
RBFNetwork	0.781	78.17%	0.803	80.36%	maxIts -1, numClusters 2, ridge 1.0E-8, clusteringSeed 1, minStdDev 2 (for Valence) clusteringSeed 3, minStdDev 0.1 (for Arousal)

VII. DISCUSSION

In the present study, we investigated the possibility of video emotion recognition using EEG-video feature fusion. Our results demonstrated that this method is effective compared to using single modality features, which improves the model performance markedly. This finding is significant because it offers a potential method of utilizing physiology-multimedia feature fusion for human visual scene emotion recognition. In this experiment, 39 participants were invited for data collection, and the number of participants is competitive among the existing studies (see Tables 1, 2 and 3). To our knowledge, the classification accuracy is also desirable. In the present study, the model performance is enhanced significantly using the multimodal method compared to using single modality, which manifests the feasibility of this method.

Hitherto, the overall classification accuracy in this research field is not satisfying, with the reported rate ranging from 40% to 90% (primarily 50%∼70%). According to the related studies, the outcome is differential and might be owing to experiment performed with different feature types, algorithms, stimuli, participants and procedures. Our experimental results of single modality are consistent with the corresponding modality recognition studies, with the accuracy ranging from 50% to 70% (see Table 3). Meanwhile, multimodal data fusion with physiological features and multimedia features is adopted increasingly by researchers, and the emotion recognition rate has generally improved in terms of classification in recent years. It is shown that many of them have achieved better performance than single modality (see Table 1 − 3). In this study, we achieved an accuracy of 96.79% for valence and 97.79% for arousal based on EEG and video feature fusion, which also indicates the conclusion of the existing studies. A detailed feature set applied in this study and the model parameters setting of each algorithm is presented herein. To further improve the efficiency, the exact mechanism for multimodal features selection and data fusion should be investigated systematically.

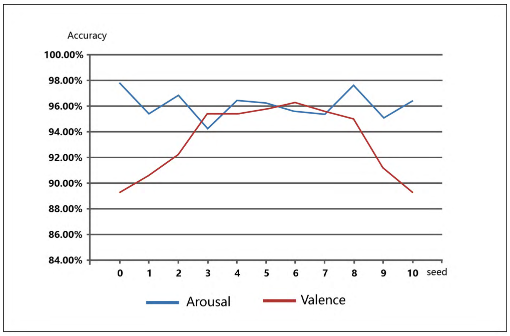


FIGURE 9. The classification accuracy for valence and arousal for different *seed*.

In most of the studies, arousal has the higher recognition rate than valence. This may be owing to the subjectivity of annotation in valence, while the annotators' perception is more consistent in arousal labeling. This phenomenon also exists widely existed in related studies. The improvement and new assumption of the valence recognition method can be a research focus in future studies.

However, we also discovered that participants sometimes had different perceptions on the same video, which might be due to the difference of personal experience or video content cognition. In this experiment, the video clips with inconsistent annotations will be abandoned for it might trigger distinct emotions. But in the real world, it is common to stimulate

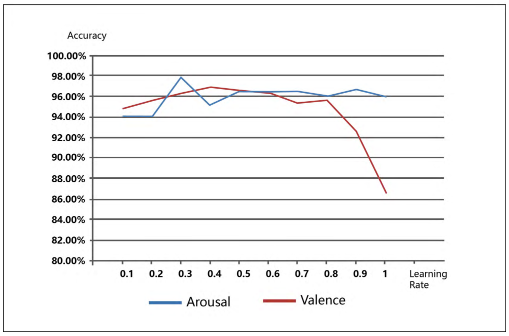


FIGURE 10. The classification accuracy for valence and arousal for different learning rate.

different video affective perceptions among people. Therefore, a user-independent multimedia emotion recognition method can be explored in further study. Automatic emotion annotation via EEG signal analysis and affective computing in various culture backgrounds can provide useful ways to solve this open research issue.

In this study, we demonstrated that video emotion recognition can be modeled with the EEG-video data fusion method. This suggests the feasibility of using the MLP method on the fusion database. The video emotion recognition model would be a valuable tool in the application of video emotion recommendation, physiology and medical research.

VIII. CONCLUSION AND FUTURE WORK

Videos are inspiring and engaging, and they could provide a fluctuating and wide range of emotion influence on its audiences. This study investigated the effect of video on human perception and the automatic classification of emotion from multimedia data and human EEG feature fusion. In addition, this study was conducted to demonstrate the effectiveness of using EEG and multimedia feature fusion to recognize video emotion by considering the video emotion expression and the individual emotion perception difference. The advantage of EEG and multimedia feature fusion and how different feature modalities could contribute to the optimal recognition results were discussed. Furthermore, the MLP algorithm was confirmed to be the best model for video emotion recognition based on the fusion database.

Herein, we used MLP to classify the video emotion based on a fusion dataset and achieved good results. However, some aspects could be further improved, which are listed below:

(1) The dataset was fairly small and could be expanded; the integrated algorithm could be applied to improve the recognition rate;

(2) Because deep convolutional neural networks (CNNs) are the state-of-art models in many multimedia analysis tasks with superior performance. DeepCNN could be applied to extract stronger feature set and improve the recognition accuracy;

(3) Different modalities of physiological features could be considered for modeling in future research, such as facial expression;

(4) A user-independent video emotion recognition method could be explored for individual video emotion retrieval.

We discovered that it was possible to recognize emotion from EEG signals in response to video scene and audio. Along with the increasing ability of computers in multimedia emotion perception, the relation between video feature and human EEG signals evoked by video content could be an important cue in emotional video retrieval.

REFERENCES

[1] R. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[2] Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen, "Affective video content analysis: A multidisciplinary insight," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 396–409, Oct./Dec. 2018.

[3] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 92–105, Apr. 2011.

[4] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[5] R. D. Fonnegra and G. M. Díaz, "Deep learning based video spatio-temporal modeling for emotion recognition," in *Proc. Int. Conf. Hum.-Comput. Interact.*, 2018, pp. 397–408.

[6] H.-B. Kang, "Affective content detection using HMMs," in *Proc. 11th ACM Int. Conf. Multimedia*, Berkeley, CA, USA, 2003, pp. 259–262.

[7] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 689–704, Jun. 2006.

[8] K. Sun and J. Yu, "Video affective content representation and recognition using video affective tree and hidden Markov models," in *Proc. 2nd Int. Conf. Affect. Comput. Intell. Interact.*, Sep. 2007, pp. 594–605.

[9] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 523–535, Oct. 2010.

[10] S. Zhang, Q. Tian, S. Jiasng, Q. Huang, and W. Gao, "Affective MTV analysis based on arousal and valence features," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun. 2008, pp. 1369–1372.

[11] M. Soleymani, J. J. M. Kierkels, G. Chanel, and T. Pun, "A Bayesian framework for video affective representation," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Oct. 2009, pp. 1–7.

[12] Y. Cui, J. S. Jin, S. Zhang, S. Luo, and Q. Tian, "Music video affective understanding using feature importance analysis," in *Proc. ACM Int. Conf. Image Video Retr.*, New York, NY, USA, Jan. 2010, pp. 213–219.

[13] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proc. 16th ACM Int. Conf. Multimedia*, Jan. 2008, pp. 677–680.

[14] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2011, pp. 2376–2379.

[15] S. E. Kahou *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. 15th ACM Int. Conf. Multimodal Interact.*, 2013, pp. 543–550.

[16] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in *Proc. 20th Int. Conf. Multimedia Modeling*, 2014, pp. 303–314.

[17] M. Gogate, A. Adeel, and A. Hussain, "A novel brain-inspired compression-based optimised multimodal fusion for emotion recognition," in *Proc. IEEE Symp. Series Comput. Intell.*, Nov./Dec. 2017, pp. 1–7. doi: 10.1109/SSCI.2017.8285377.

[18] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 255–270, Apr./Jun. 2018.

[19] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *Proc. IEEE 10th Int Symp. Multimedia*, Dec. 2008, pp. 228–235.

[20] H.-B. Kang, "Affective contents retrieval from video with relevance feedback," in *Proc. 6th Int. Conf. Asian Digit. Libraries*, Kuala Lumpur, Malaysia, Dec. 2003.

[21] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, vol. 7, no. 1, pp. 143–154, Feb. 2005.

[22] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, "Affective visualization and retrieval for music video," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 510–522, Oct. 2010.

[23] S. Wang, H. Lin, and Y. Hu, "Affective classification in video based on semi-supervised learning," in *Advances in Neural Networks*. 2011, pp. 238–245.

[24] Y. Matsuda, D. Fedotov, Y. Takahashi, Y. Arakawa, K. Yasumoto, and W. Minker, "EmoTour: Estimating emotion and satisfaction of users based on behavioral cues and audiovisual data," *Sensors*, vol. 18, no. 11, p. 3978, 2018.

[25] J. Niu, X. Zhao, L. Zhu, and H. Li, "Affivir: An affect-based Internet video recommendation system," *Neurocomputing*, vol. 120, pp. 422–433, Nov. 2013.

[26] S. Mo, J. Niu, Y. Su, and S. K. Das, "A novel feature set for video emotion recognition," *Neurocomputing*, vol. 291, pp. 11–20, May 2018.

[27] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 410–430, Oct./Dec. 2015.

[28] Y.-P. Lin *et al.*, "EEG-based emotion recognition in music listening," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1798–1806, Jul. 2010.

[29] D. Nie, X.-W. Wang, L.-C. Shi, and B.-L. Lu, "EEG-based emotion recognition during watching movies," in *Proc. IEEE/EMBS 5th Int. Conf. Neural Eng.*, Apr./May 2011, pp. 667–670.

[30] X. W. Wang, D. Nie, and B. L. Lu, "Emotional state classification from EEG data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94–106, Apr. 2014.

[31] Z. Lan, O. Sourina, L. Wang, and Y. Liu, "Stability of features in real-time EEG-based emotion recognition algorithm," in *Proc. Int. Conf. Cyberworlds*, Oct. 2014, pp. 137–144.

[32] J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," *Expert Syst. Appl.*, vol. 47, pp. 35–41, Apr. 2016.

[33] A. M. Bhatti, M. Majid, S. M. Anwar, and B. Khan, "Human emotion recognition and analysis in response to audio music using brain signals," *Comput. Hum. Behav.*, vol. 65, pp. 267–275, Dec. 2016.

[34] E. Kroupi, J.-M. Vesin, and T. Ebrahimi, "Subject-independent odor pleasantness classification using brain and peripheral signals," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 422–434, Oct./Dec. 2016.

[35] C. Shahnaz, Shoaib-Bin-Masud, and S. M. S. Hasan, "Emotion recognition based on wavelet analysis of Empirical Mode Decomposed EEG signals responsive to music videos," in *Proc. Region 10 Conf.*, Nov. 2017, pp. 424–427.

[36] Y. Li, W. Zheng, Z. Cui, Y. Zong, and S. Ge, "EEG emotion recognition based on graph regularized sparse linear regression," *Neural Process. Lett.*, vol. 49, no. 2, pp. 555–571, 2018.

[37] Y. Y. Lee and S. Hsieh, "Classifying different emotional states by means of EEG-based functional connectivity patterns," *PLoS ONE*, vol. 9, no. 4, 2014, Art. no. e95415.

[38] S. Liu, J. Tong, M. Xu, J. Yang, H. Qi, and D. Ming, "Improve the generalization of emotional classifiers across time by using training samples from different days," in *Proc. IEEE EMBS*, Aug. 2016, pp. 841–844.

[39] N. Ran, "EEG-based emotion recognition," M.S. thesis, Dept. Comput. Sci., School Electron. Electr. Eng., Shanghai Jiao Tong Univ., Shanghai, China, 2012.

[40] Z. Lin, "A Theoretical research on EEG-based emotional tagging of music videos," M.S. thesis, School Inf. Sci. Eng., Lanzhou Univ., Lanzhou, China, 2014.

[41] A. Singhal, P. Kumar, R. Saini, P. P. Roy, D. P. Dogra, and B.-G. Kim, "Summarization of videos by analyzing affective state of the user through crowdsource," *Cogn. Syst. Res.*, vol. 52, pp. 917–930, Dec. 2018.

[42] K. Barjinder, S. Dinesh, and P. R. Partha, "EEG based emotion classification mechanism in BCI," *Procedia Comput. Sci.*, vol. 132, pp. 752–758, Jan. 2018.

[43] I. S. Moon and S. Mandeep, "Development of a real time emotion classifier based on evoked EEG," *Biocyernetics Biomed. Eng.*, vol. 37, no. 3, pp. 498–509, 2017.

[44] J. Moon, Y. Kwon, J. Park, and W. Yoon, "Detecting user attention to video segments using interval EEG features," *Expert Syst. Appl.*, vol. 115, pp. 578–592, Jan. 2019.

[45] Y. Zhang, X. Ji, and S. Zhang, "An approach to EEG-based emotion recognition using combined feature extraction method," *Neurosci. Lett.*, vol. 633, pp. 152–157, Oct. 2016.

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

[47] K. P. Wagh and K. Vasanth, "Electroencephalograph (EEG) based emotion recognition system: A review," in *Innovations in Electronics and Communication Engineering* (Lecture Notes in Networks and Systems). Singapore: Springer, 2019, pp. 37–59.

[48] L. R. Christensen and M. A. Abdullah, "EEG emotion detection review," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol.*, May/Jun. 2018, pp. 1–7.

[49] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affect. Comput.*, to be published. doi: 10.1109/TAFFC.2017.2714671.

[50] J.-L. Hsu, Y.-L. Zhen, T.-C. Lin, and Y.-S. Chiu, "Affective content analysis of music emotion through EEG," *Multimedia Syst.*, vol. 24, no. 2, pp. 195–210, Mar. 2018.

[51] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 209–222, Jul. 2015.

[52] K. Takahashi, "Remarks on SVM-based emotion recognition from multimodal bio-potential signals," in *Proc. 13th IEEE Int. Workshop Robot Hum. Interact. Commun.*, Sep. 2003, pp. 95–100.

[53] S. Koelstra *et al.*, "Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos," in *Proc. Int. Conf. Brain Inform.*, 2010, pp. 89–100.

[54] S. Koelstra *et al.*, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan./Mar. 2012.

[55] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[56] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, Apr. 2012.

[57] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, Jan./Mar. 2016.

[58] M. B. H. Wiem and Z. Lachiri, "Emotion classification in arousal valence model using MAHNOB-HCI database," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 3, pp. 318–323, 2017.

[59] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS)," *IEEE Access*, vol. 7, pp. 57–67, 2018.

[60] L. Duan, H. Ge, Z. Yang, and J. Chen, "Multimodal fusion using kernel-based ELM for video emotion recognition," *Proc. ELM*, vol. 1, pp. 371–381, Jan. 2016.

[61] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "MPED: A multi-modal physiological emotion database for discrete emotion recognition," *IEEE Access*, vol. 7, pp. 12177–12191, 2019.

[62] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *Proc. 2nd ACM Workshop Multimedia Semantics*, 2008, pp. 32–39.

[63] B. Voytek and R. T. Knight, "Prefrontal cortex and basal ganglia contributions to visual working memory," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 42, pp. 18167–18172, 2010.

[64] H. Bakardjiana, T. Tanakaa, and A. Cichockia, "Emotional faces boost up steady-state visual responses for brain-computer interface," *NeuroReport*, vol. 22, pp. 121–125, 2011.

[65] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Therapy Experim. Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[66] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, pp. 9–21, Mar. 2004.

[67] M. H. J. Gruber, "Statistical digital signal processing and modeling," *Technometrics* , vol. 39, no. 3, pp. 335–336, 1996.

[68] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affective Comput.*, vol. 7, no. 2, pp. 190–202, Apr./Jun. 2015.

[69] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

[70] N. Midha and V. Singh, "Classification of E-commerce products using RepTree and K-means hybrid approach," in *Big Data Analytics* (Advances in Intelligent Systems and Computing). Singapore: Springer, 2018, pp. 265–273.

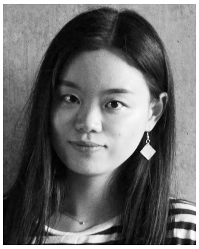
[71] S.-H. Wang, H.-T. Li, E.-J. Chang, and A.-Y. Wu, "Entropy-assisted emotion recognition of valence and arousal using XGBoost classifier," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.*, 2018, pp. 249–260.

[72] R. Collobert and S. Bengio, "Links between Perceptrons, MLPs and SVMs," in *Proc. 21st Int. Conf. Mach. Learn.*, Sep. 2004, p. 23.

[73] M. Drmota, *Random Trees*. Vienna, Austria: Springer, 2009.

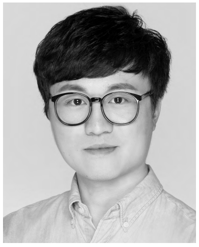
[74] W. Chen, X. Yan, Z. Zhao, H. Hong, D. T. Bui, and B. Pradhan, "Spatial prediction of landslide susceptibility using data mining-based kernel logistic regression, naive Bayes and RBFNetwork models for the Long County area (China)," *Bull. Eng. Geol. Environ.*, vol. 78, no. 1, pp. 247–266, 2018.

BAIXI XING received the Ph.D. degree in digital art and design from Zhejiang University, Hangzhou, China, in 2014. She is currently an Assistant Professor with the School of Computer Science and Technology, Hangzhou Dianzi University. She is currently holding a postdoctoral position with the College of Computer Science and Technology, Zhejiang University. Her research interests include affective computing and multimedia retrieval. She is currently focusing in multimodal emotion recognition and cross-media retrieval. She is also interested in the research of human–computer interaction and user experience design.



HUI ZHANG received the B.Eng. degree in computer science and technology from Zhejiang University, China, in 2016, where she is currently pursuing the Ph.D. degree in digital art and design with the College of Computer Science and Technology. Her research interests include music information retrieval, affective computing, and human–computer interaction. She is a Student Member of the China Computer Federation and Association for Computing Machinery.

KEJUN ZHANG received the Ph.D. degree in computer science from Zhejiang University, China, where he is currently an Associate Professor with the College of Computer Science. His research interests include music information retrieval, bioinformatics, evolutionary computation, and machine learning. He has published many research papers in various reputable journals and conference proceedings.



LEKAI ZHANG received the Ph.D. degree in digital art and design from the College of Computer Science and Technology, Zhejiang University, China. He is fully indulged in affective computing, design thinking and user-centered design, and carrying out integrated interactive works. His research interests include the field of ubiquitous computing, human–computer interaction, interactive user interfaces, and multimedia interaction.

XINDA WU is currently pursuing the bachelor's degree with the School of Computer Science and Technology, Hangzhou Dianzi University. His research interests include the field of affective computing, human–computer interaction, and multimedia information retrieval.

XIAOYING SHI received the Ph.D. degree in control theory and control engineering from the Zhejiang University of Technology, China, in 2014. She is currently an Assistant Professor with the School of Computer Science and Technology, Hangzhou Dianzi University. Her research interests are in visual analysis, machine learning, and urban transportation.

SHANGHAI YU is currently pursuing the bachelor's degree with the School of Computer Science and Technology, Hangzhou Dianzi University. His research interests include the field of affective computing, speech emotion analysis, and video sentiment analysis.

SANYUAN ZHANG received the bachelor's, master's, and Ph.D. degrees from the Department of Mathematics, Zhejiang University, Hangzhou, China, in 1986, 1989, and 1992, respectively, where he is currently a Professor with the College of Computer Science and Technology. His research interests include image and video processing and computer graphics.