UNIVERSITY
OF TURKU

A Dual-Modality Emotion Recognition System of EEG and Facial Images and its Application in Educational Scene

Robotics and Autonomous Systems
Master's Degree Programme in Information and Communication Technology
Department of Computing, Faculty of Technology
Master of Science in Technology Thesis

Author:
Ruijin Li

Supervisors:
MSc (Tech) Jorge Peña Queralta
Assoc. Prof. Tomi Westerlund

December 2022

Master of Science in Technology Thesis
Department of Computing, Faculty of Technology
University of Turku

Subject: Robotics and Autonomous Systems
Programme: Master’s Degree Programme in Information and Communication Technology
Author: Ruijin Li
Title: A Dual-Modality Emotion Recognition System of EEG and Facial Images and its Application in Educational Scene
Number of pages: 69 pages
Date: December 2022

Abstract.

With the development of computer science, people's interactions with computers or through computers have become more frequent. Some human-computer interactions or human-to-human interactions that are often seen in daily life: online chat, online banking services, facial recognition functions, etc. Only through text messaging, however, can the effect of information transfer be reduced to around 30% of the original. Communication becomes truly efficient when we can see one other's reactions and feel each other's emotions.

This issue is especially noticeable in the educational field. Offline teaching is a classic teaching style in which teachers may determine a student's present emotional state based on their expressions and alter teaching methods accordingly. With the advancement of computers and the impact of Covid-19, an increasing number of schools and educational institutions are exploring employing online or video-based instruction. In such circumstances, it is difficult for teachers to get feedback from students. Therefore, an emotion recognition method is proposed in this thesis that can be used for educational scenarios, which can help teachers quantify the emotional state of students in class and be used to guide teachers in exploring or adjusting teaching methods.

Text, physiological signals, gestures, facial photographs, and other data types are commonly used for emotion recognition. Data collection for facial images emotion recognition is particularly convenient and fast among them, although there is a problem that people may subjectively conceal true emotions, resulting in inaccurate recognition results. Emotion recognition based on EEG waves can compensate for this drawback. Taking into account the aforementioned issues, this thesis first employs the SVM-PCA to classify emotions in EEG data, then employs the deep-CNN to classify the emotions of the subject's facial images. Finally, the D-S evidence theory is used for fusing and analyzing the two classification results and obtains the final emotion recognition accuracy of 92%. The specific research content of this thesis is as follows:

1) The background of emotion recognition systems used in teaching scenarios is discussed, as well as the use of various single modality systems for emotion recognition.

2) Detailed analysis of EEG emotion recognition based on SVM. The theory of EEG signal generation, frequency band characteristics, and emotional dimensions is introduced. The EEG signal is first filtered and processed with artifact removal. The processed EEG signal is then used for feature extraction using wavelet transforms. It is finally fed into the proposed SVM-PCA for emotion recognition and the accuracy is 64%.

3) Using the proposed deep-CNN to recognize emotions in facial images. Firstly, the Adaboost algorithm is used to detect and intercept the face area in the image, and the gray level balance is performed on the captured image. Then the preprocessed images are trained and tested using the deep-CNN, and the average accuracy is 88%.

4) Fusion method based on decision-making layer. The data fusion at the decision level is carried out with the results of EEG emotion recognition and facial expression emotion recognition. The final dual-modality emotion recognition results and system accuracy of 92% are obtained using D-S evidence theory.

5) The dual-modality emotion recognition system's data collection approach is designed. Based on the process, the actual data in the educational scene is collected and analyzed. The final accuracy of the dual-modality system is 82%. Teachers can use the emotion recognition results as a guide and reference to improve their teaching efficacy.

Key words: emotion recognition, EEG signal, facial image, SVM, CNN, education

Table of Contents

1 Introduction

1.1 Research Background and Significance

People’s life is always accompanied by emotions, and the most common emotions are joy, anger, sadness, and so on. People have emotions and show their feelings in their daily encounters. Emotions are inextricably linked to life since they affect not only the individual but also the others around him in subtle ways. There are many disciplines related to emotion[1], such as medicine, psychology, neuroscience, sociology[1], etc. Emotion research is a relatively abstract and complex subject. Emotion has an impact on every element of a person's life, including cognition, decision-making, and action. When communicating with people, both sides can respond differently by observing each other’s emotional changes. The same is true for AI products. Speech recognition, face recognition, and data analysis are all areas where AI is now being employed. However, these devices are unable to detect human emotions during human-computer interaction, resulting in an insufficiently gratifying interactive experience. The computer must learn the user's emotions in order to establish a more natural human-computer interface, and this is also a future development trend for AI products.

Similarly, the implementation of emotion recognition in the educational system is critical[2]. Students with positive emotions perform better, learn faster and more easily than students with negative emotions. Teachers who understand how different teaching methods affect students' emotions can make timely adjustments, locate the most successful way, and increase students' excitement for learning.

People's emotional interactions are complicated and recognizing emotions using only one modality offers advantages and disadvantages. From the perspective of multimodality, effective information from different modalities can be extracted to improve the accuracy of the emotion recognition model[3]. With the development of AI technologies such as speech recognition, natural language processing, and computer vision, it has become possible to apply these technologies to practical scenarios of emotion recognition.

When a person’s emotions change, their body changes accordingly, and we can see the changes from different aspects. Researchers can study emotion recognition in a variety of ways, including the following:

By facial expressions[4]: Facial expression is an intuitive mirror of human emotions, and expression recognition has always been one of the important research topics in computer vision. The facial expression recognition is based on face detection. Deep learning technology is frequently utilized in emotion recognition. It has a high degree of accuracy and strong robustness and can be applied in teaching scenarios. When it comes to teaching, the face detection system can reflect how involved the students are, evaluate the learning effectiveness of students and help teachers adjust teaching methods.

By texts[5]: Since people more often use text messages for communication these days, emotion can be extracted from the texts and be analyzed for the author's subjective sentiments[6]. First of all, it is necessary to establish an emotional feature thesaurus to store emotion feature words and attribute information. Then segment the text using the word segmentation tool. The processes word segments are matched with the emotion features. Finally, the emotional tendency of this text is determined according to the extracted word segmentation and attribute information.

By gestures[7]: In real life, people are frequently confronted with a variety of situations. Body movements are used to express emotions when there is no method to communicate verbally and when it is difficult to observe facial expressions properly. First of all, we need to establish a dataset of human motion related to emotions in daily life scenarios, and the focus of this dataset is on body movements. Through gesture capture data and image data, features such as body, force, space, and shape are extracted. Finally, emotion recognition results are obtained through analyzing gestures[8].

By speeches: Human voices hold a wealth of information, including semantic, emotional, and other types of information. Speech emotion recognition refers to automatically identifying emotions contained in the language. The reason why humans can capture the changes in an emotional state of other people by listening to them is that the human brain can perceive and understand the emotional information in speech. Automatic speech recognition includes emotion recognition, language content, etc. In real life, computers could simulate the human brain's analysis of speech signals through internal calculations. Speech recognition is also frequently utilized in everyday life, for example, in intelligent speakers, voice navigation, and safe driving reminders[9].

By physiological signals[10]: Human emotion changes are accompanied by changes in physiological signals. Physiological signals have the advantage of being able to better reflect true emotional states than facial expressions or spoken messages. EEG, eye movement, EMG, electrodermal, ECG, and respiration are the most common physiological signals employed for emotion recognition. Because the frequency of these physiological signals is usually low, and the acquisition is easily affected by the external environment, special equipment is required for physiological signal collection. To improve the quality and accuracy of emotion recognition, the signal must be further preprocessed after acquisition.

Emotion analysis on a single modality is highly advanced, but for data containing multiple modalities, the additional information makes emotion recognition more challenging. At the same time, in the fields of single modality emotion recognition, one type of dataset can be easily affected by various noises, such as easily disguised facial expressions, and it is difficult to completely portray subjects' true emotional state. To improve the accuracy of emotion recognition, some researchers consider combining two or more different data types and doing the analysis.

In this research, a new strategy for improving emotion recognition accuracy is proposed. Both facial expressions and EEG signals collected from subjects are used for emotion recognition. EEG signals can better depict participants' true feelings, which compensates for the camouflage shortcomings of emotion identification based solely on facial expressions. This emotion recognition system, which uses EEG data and facial images to recognize emotions, could be employed in the educational scene.

The emotional data of students has guiding value for teaching and research, educational institutions, and teachers themselves. Large-scale negative feedback may mean that the teaching arrangement is unreasonable. The emotional status of students participating in the experiment can be collected and analyzed by teachers. Teachers can enhance their teaching methods and learning efficiency based on the results of emotion recognition.

1.2 Literature Review

1.2.1 Emotion Evocation

At present, most of the experiments related to emotion recognition are carried out in a controllable environment. An important precondition in emotion recognition experiments is how to evoke different emotions of subjects[11]. Picard[12] divided the emotion evocation method into two types, one is subject evocation, and the other one is event evocation. Subject evocation refers to that subjects need to recall an event from their life that can remind them of a certain emotion. Even evocation refers to a method of evoking subjects' specific emotions through stimulus content such as text, pictures, video clips, audio, etc. Although the subject evocation approach can effectively evoke the desired emotion, it does so only with the participants' conscious involvement. It's possible that it'll result in an unmanageable experimental situation. More researchers are choosing to use the event evocation approach and conduct the emotion evocation experiment in order to keep the experiment under control and reduce other distractions.

The event evocation method refers to presenting materials with emotional tendencies to the subjects, thereby evoking corresponding emotions of the subjects. Based on the different kinds of materials used in the experiment, they can be mainly divided into visual stimulation materials, auditory stimulation materials, and olfactory stimulation materials. Multi-channel emotion evocation is a technique for bettering emotion evocation by combining visual, aural, olfactory, and other evocation components. In the event evocation method, the use of video clips as stimulation materials not only synthesizes the advantages of auditory sense and visual sense but also effectively evokes emotions. That's the reason why video clips are widely used in the real emotion evocation experiment.

1.2.2 Literature Review Based on EEG Signals

Emotional physiological signals are created spontaneously by the human neurological and endocrine systems and are not easily influenced by external factors[13]. There are two common methods of emotion recognition based on physiological signals. The first one is emotion recognition based on the autonomic nervous system, and the data types measured mainly include external physiological manifestations such as human respiratory rate, skin electrical signals, and heart rate. The emotion classification result is obtained through the analysis of these data. Picard et al. of the Massachusetts Institute of Technology in the United States identified eight different emotions of calm, anger, disgust, sadness, pleasure, romance, joy, and fear through measurement and analysis of the human autonomic nervous system. Despite the fact that the physiological signals of these autonomic nerve systems cannot be disguised and can obtain true data, they are unsuitable for practical applications due to their low precision and lack of adequate evaluation standards[15].

The second method is emotion recognition based on the central nervous system. The central nervous system refers to the brain electrical activity. The body's central nervous system responds differently in different emotional states. Emotion classification results can be obtained through the processing of EEG signals and feature extraction. Common central nervous system-based recognition methods include fMRI and EEG. Due to the large size and high price of fMRI equipment, it is not suitable for practical application. Therefore, emotion recognition based on EEG is currently a more commonly used method.

EEG[15] is a spatially discrete, non-stationary random signal produced by the central nervous system which can directly record changes in subjects' scalp potential. For emotion recognition, traditional research methods often extract linear and nonlinear features from EEG signals, ignoring the interplay between brain areas. In recent years, more researchers have introduced complex network theory into the study of EEG emotion recognition and explored the mechanism of emotion generation by building a brain function network then doing the emotion recognition.

The main steps in emotion recognition using EEG signals[16] include EEG signal acquisition, data preprocessing, feature extraction, and emotion recognition. The acquisition of EEG signals is generally achieved by placing physical electrodes on the scalp. The number of electrodes changes along with different acquisition devices. The most commonly used are 16 electrodes, 32 electrodes, and 64 electrodes. These electrodes are placed in the right position on the scalp of the subjects' brains according to the 10-20 system[17] electrode placement method and are used to collect EEG signals generated by different brain regions.

In the process of collecting EEG signals, some interference noise will be collected due to the influence of the external environment, eye movements, muscle movements, and other factors. Data preprocessing is to remove the interference noise in the original EEG signals and obtain purer EEG signals. The common preprocessing method includes filtering, PCA, independent component analysis, etc. Bartels et al. combined blind signal separation, independent component analysis, and SVM to propose an effective artifacts removal processing method. By analyzing the characteristics of each algorithm, the ocular artifacts are removed using the Amuse algorithm in BBS. The infomax algorithm in the independent component analysis is used to remove the EMG signals. Figure 1-1 and Figure 1-2 show a good result in artifacts removal.

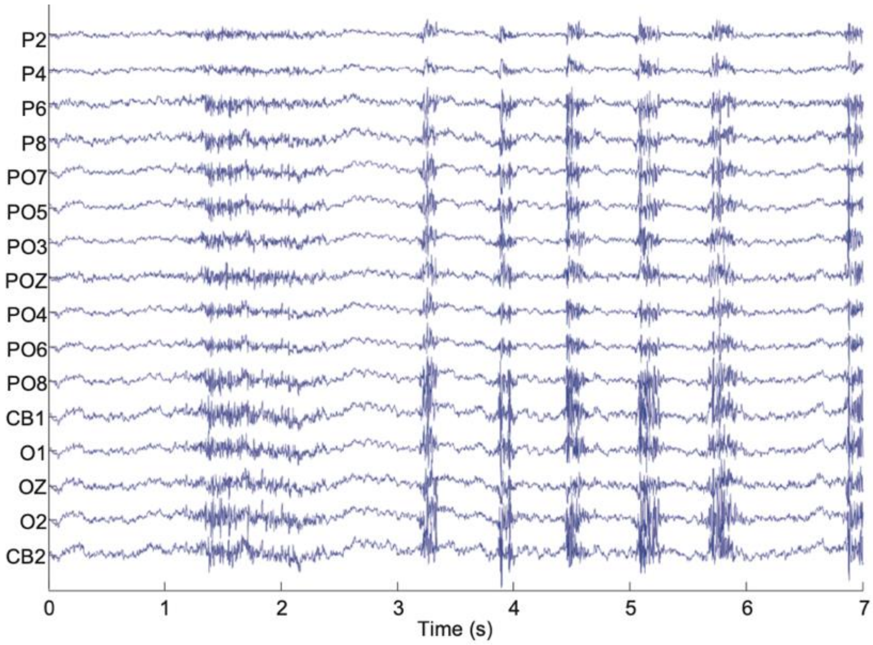


Figure 1-1 Raw EEG recordings with strong artifacts[19]

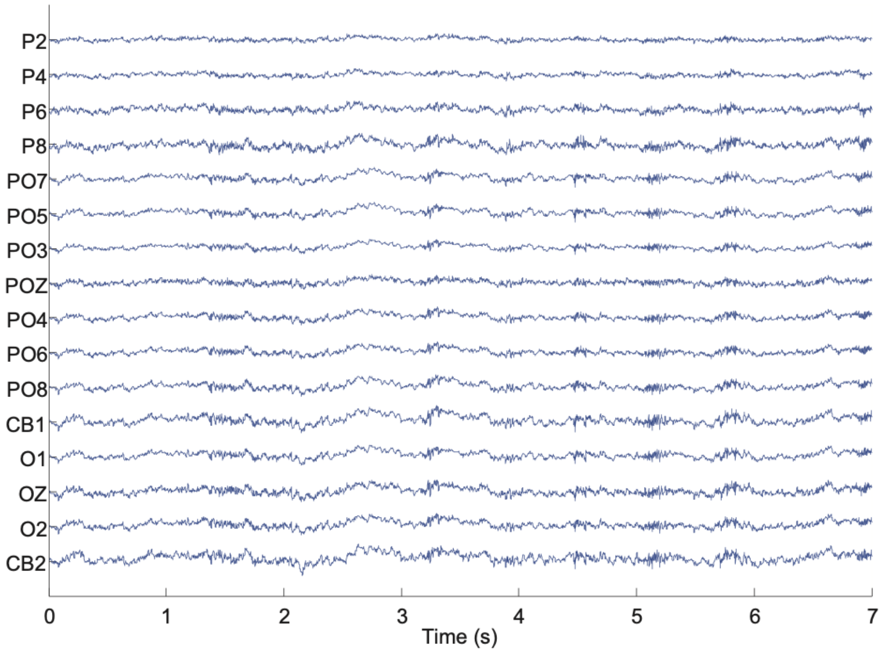


Figure 1-2 Processes EEG after artifacts removal[19]

The feature extraction[19] of EEG signal is to use EEG as the source signal to determine various parameters and use parameters as a vector to characterize the EEG features. Its purpose is to reduce the dimensionality of the EEG, highlight emotional characteristics, and be used to study the emotional state of subjects. Choosing features will directly affect the accuracy and relevance of the emotion recognition model. Usually, the EEG signal features used are time-domain features, frequency features, time-frequency features, and nonlinear features. Time-domain features are the most intuitive and easily obtained in several types of features. This method often uses the information in the time domain after removing artifacts as features or uses signal statistics in the time domain as features. Frequency features are mainly based on the waveform of each frequency band of the EEG signal. It is incomplete to consider only the frequency domain characteristics of the EEG signal and ignore the time-domain features. As a result, more and more researchers begin to combine the time and frequency domains to find EEG features which reflect both the time and frequency features.

Recognition is identifying EEG patterns corresponding to different emotional states by extracting EEG features and then classifying them. Common recognition methods[20] are unsupervised learning, semi-supervised learning, and supervised learning. Some classifiers such as self-organizing map, SVM, K-nearest neighbor are used for emotion recognition. One of the goals of emotion recognition using EEG is to select features that represent emotional states. Through the optimization model, the classification accuracy of emotional states is improved as much as possible, which provides a reliable guarantee for the application of EEG in the field of emotion recognition. Another goal is to identify the brain regions and frequency bands most associated with emotional states in the research process, providing a physiological basis for the application of EEG in the field of emotion recognition.

1.2.3 Literature Review Based on Facial Expressions

Facial expressions are a very important form of communication. In 1971, Ekman et al.[21] for the first time divided expressions into six basic types, sad, happy, fear, disgust, surprise, and angry. With the development and deepening of facial expression research, FACS based on AU is proposed, which explains the connection between AU and expressions by analyzing the characteristics of AU. A facial image contains a lot of information and the expressions at different moments in the time sequence are not the same. It is necessary to extract effective information such as texture features and facial features in the image when recognizing emotions. The extraction of this effective information is of great significance for the improvement of emotion recognition speed and accuracy. The robustness and integrity of the extracted features will have a decisive impact on the final recognition results.

The traditional facial emotion recognition method[22] includes four basic steps: raw facial image input, data preprocessing, feature extraction, and emotion recognition. Among those steps, feature extraction is the most important step in the traditional method, which requires meeting the final emotion recognition needs. These features are extracted artificially, then entered into the classifier. For facial images, they are usually collected in the laboratory, which will then be converted to grayscale images and normalized. Most noise will be removed by feature extraction and images finally are fed into the classifier to obtain the emotion recognition result. The facial emotion recognition process based on deep learning is similar to the traditional method, however, in deep learning, there is no need to artificially select features. It is done automatically by the deep learning network. The deep learning method requires a large amount of data. If the amount of data is too small in the training set, there may be overfitting, resulting in poor generalization performance. Therefore, translation, rotation, cutting, adding noise always are used for data enhancement. The commonly used deep neural network model is CNN, which includes convolutional layers, pooling layers, and connected layers. Some networks combine CNN with LSTM, and the time sequence features are obtained from image features. Although more and more expression datasets have been public in recent years, the shortcomings of small data size and unbalanced data types are still common problems for emotion recognition. As deep learning becomes more widely applied in facial expression recognition, training neural networks require richer, more balanced datasets. In response to this problem, researchers begin to augment the original data set using data-enhancing methods. Initially, geometric transformations of original images is a common method. Simard et al.[23] proposed increasing the sample size by panning, rotating, and skewing. Lopes et al.[24] choose to add random noise near the eyes of the original image by using a two-dimensional Gaussian distribution, and then maintain the balance of the two eye positions by rotating the operation, resulting in a new sample image. Krizhevsky et al.[25] randomly crops the original image to form a fixed-size subsample, and then flipped each subsample horizontally, which expands the training set. Adjusting the brightness and contrast of images is also a common data enhancement method. Adjusting the brightness of original images, can not only expand the sample size but also weaken the impact of brightness problems on emotion recognition.

At present, the research of facial expression recognition has made great progress. However, there are still some problems with facial expression recognition[26], mainly in the following aspects:

1) The types of expressions are not enough. The expressions commonly used in daily life are not limited to these six basic expressions. Therefore, it may be possible to have a low recognition accuracy for complex expressions.

2) The traditional feature extraction and classification methods generally verify the accuracy through standard facial expression datasets. It may lead to insufficient consideration for complex and changeable environments, weakening the applicability of the emotion recognition system.

3) Traditional feature extraction methods are difficult to extract deeply hidden features in human facial expressions. Deep learning algorithms such as CNN can extract features that are difficult for a human to consider, but training complex neural network requires a lot of computation and time.

4) Facial expressions are easy to observe and be used for classification, but they are also easy to disguise. For example, if the subject chooses to pretend the facial expression to cover up his true emotions, then the final emotion classification result is not a good reference.

1.3 Research Purposes

Teachers want to implement more effective teaching approaches to boost students' enthusiasm for learning, whether through traditional teaching or online distance learning. Students' emotional states have a significant impact on their ability to learn, so their emotional states are a crucial criterion for determining whether or not a teaching approach is effective. Using various approaches to collect emotional feedback from students might help the teacher change their teaching methods during actual teaching activities.

EEG data and human facial expressions are both representations of a person's emotional state. Considering combine the EEG emotion recognition and facial image emotion recognition, the advantages of both ways could be obtained. In this way, using the dual-modality emotion recognition system could increase accuracy, comparing with the single modality system. In this thesis, we focus on the dual-modality emotion recognition model based on EEG signals and facial expressions that can be applied in educational scene.

1.4 Innovation Points

The main work and innovation points of this thesis include the following aspects:

1) A dual-modality emotion recognition model based on EEG data and facial expressions is suggested to address the problem of single modality emotion recognition's inaccuracy. Using the dual-modality system, the shortcomings of a single modality system can be compensated. The dual-modality system proposed in the thesis can combine the advantages of two emotion recognition systems to improve classification accuracy.

2) The proposed dual-modality emotion recognition system based on EEG signals and facial expressions is applied to practical educational scenes. This methodology can be used to help teachers change their teaching methods by serving as a reference for curriculum creation. The "quantification" of the educational process is achieved by the emotion recognition system. Although teachers can see the actual state of students in class, it is difficult to form effective and instructive data after class, and the data of students' learning status is also missing. For educational evaluation and instructional advice, this dual-modality system has specific reference and guiding relevance.

1.5 Thesis Structure

This thesis is divided into six chapters, which are arranged as follows:

The first chapter, Introduction: focuses on the research background and significance of the thesis's core issue, as well as the emotion recognition algorithms that are based on various forms of data. It mostly covers the background and substance of EEG-based emotion detection and facial expression-based emotion recognition. Following that, the thesis's major goals and related innovative points are discussed.

The second chapter, Emotion recognition based on EEG signals: the first part describes the main steps for emotion recognition based on EEG signals, as well as the theory of EEG signal generation, characteristics, and emotional dimensions. The second part emphasizes the five

EEG frequency bands. The data set used in this thesis, which is an EEG public dataset with video data, is detailed in the third section of this chapter. The process of preprocessing EEG signals is described in the fourth part. The first stage in preprocessing is filtering, the second step is manual denoising, and the third step is ocular artifact removal using the BBS method. The fifth part explains how wavelet transformations are used to extract EEG features, focusing on the wavelet transform principle and the wavelet decomposition findings. The sixth part explains how SVM is used to recognize EEG emotions, focusing on the mathematical principles of SVM. The improved algorithm SVM-PCA is proposed and the accuracy results are obtained after training the neural network on the DEAP dataset. The summary is at the end of this chapter.

The third chapter, Emotion recognition based on facial expressions: the first part of this chapter describes the basic steps for emotion recognition based on facial expressions. The second parts covers the DEAP dataset's face video preprocessing methods, which primarily involve image capture, face detection, and image cropping. The Adaboost algorithm is used for face detection in this part. The third part introduces the preprocessing methods for cropped facial images. Preprocessing methods include grayscale conversion, normalization, and grayscale equalization. The proposed deep-CNN model for emotion recognition is introduced in the fourth part. This part covers the fundamentals of CNN as well as the network structure employed in this system. The last part is the chapter summary.

The fourth chapter, Dual-modality decision-level fusion: the first part describes the basic steps for dual-modality decision level data fusion. The theory of multi-source information fusion is introduced in the second part, which mainly includes three methods of information fusion: data-level fusion, feature-level fusion, and decision-level fusion. The third part describes the D-S method used in this system which is based on decision level fusion. The final emotion recognition results of EEG signals and facial expressions are also shown in this part. The chapter summary is introduced in the final section.

The fifth chapter, Application of dual-modality emotion recognition system in the educational scene: The first part summarizes the current situation of teaching and the significance of emotion recognition in educational applications. The steps for designing experiments and collecting data are covered in the second part. The goal is to record EEG signals and facial expressions from students during the experiment for further analysis. The third part involves collecting and analyzing experimental data from the subject using the method proposed in the previous part. To assess the student's learning status, the proposed dual modality emotion recognition system is applied. It is feasible to provide teachers with recommendations on how to adjust their teaching approaches based on the evaluation findings. This chapter's summary is in the last part.

The sixth chapter: Summary and outlook: it summarizes the significance of this research and introduces the main process and ideas of this thesis. The outlooks of this dual-modality emotion identification system are provided in the last part of this chapter for future research possibilities.

2 Proposed SVM-PCA for Emotion Recognition Based on EEG Signals

2.1 Basic Steps for Emotion Recognition Based on EEG Signals

In the emotion recognition based on EEG signals, the EEG acquisition device is used in the experiment, which is placed on the subject's head to collect physiological data. After that, the EEG signals are purified using a filter, manual denoising, and ocular artifact removal. The wavelet technique is also used to extract major features from EEG recordings. The proposed SVM-PCA neural network is utilized to classify emotions and obtain the final emotion identification result after selecting features.

2.2 The Generation, Characteristics, and Emotion Theories of EEG

EEG signals are formed by the sum of many neuronal synapses in the cerebral cortex[27]. It is the result of a combination of a large number of neurons. How the next neuron acts depend on the neurotransmitters released by the nerve endings from the last neuron. Excitatory neurotransmitters raise resting potentials and increase neuronal excitability, causing excitatory postsynaptic potentials. Inhibitory neurotransmitters reduce resting potentials and cause neuronal excitability to decrease, leading to inhibitory postsynaptic potentials.

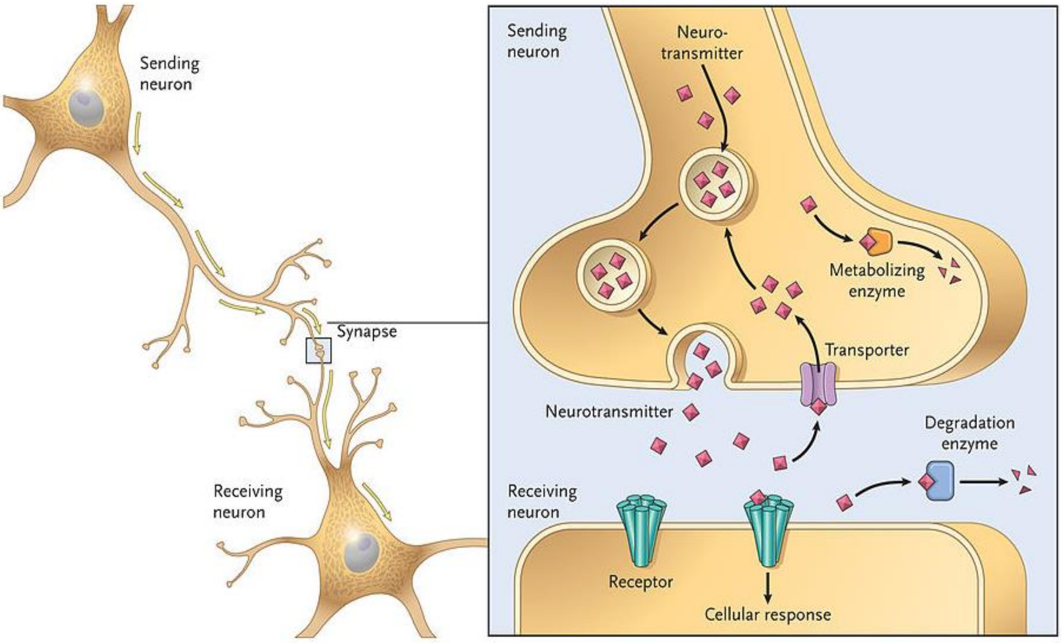


Figure 2-1 The generation mechanism of EEG signals[29]

In most physiological cases, synaptic activity is the most important component of EEG potentials. The main activities of neuronal synapses, the transmission mode of releasing neurotransmitters, and the physiological principle of EEG signal generation can be clearly understood from Figure 2-1.

EEG signals usually imply the neuronal activity of the human brain and contain information such as human physiology, emotional state[30] and thoughts. EEG signals have the following three main characteristics due to their unique acquisition method.

1. EEG signals are weak and susceptible to interference

EEG signals are collected by electrodes placed on the subject's head. The received signals are relatively weak and susceptible to interference such as muscle movements, blinking, light effects, etc. That noise is not the main research object, so it is necessary to preprocess the EEG signals. The noise should be removed and extract effective EEG features for later analysis.

2. Easy to obtain frequency domain features

Usually, EEG bands range from 0.1Hz to 100Hz[31]. In most research, EEG is divided into five frequency bands, namely Delta δ band, Theta θ band, Alpha α band, Beta β band, and Gamma γ band. These five frequency bands match with different cognitive characteristics, as shown in Table 2-1. The waveforms of five frequency bands are shown in Figure 2-2.

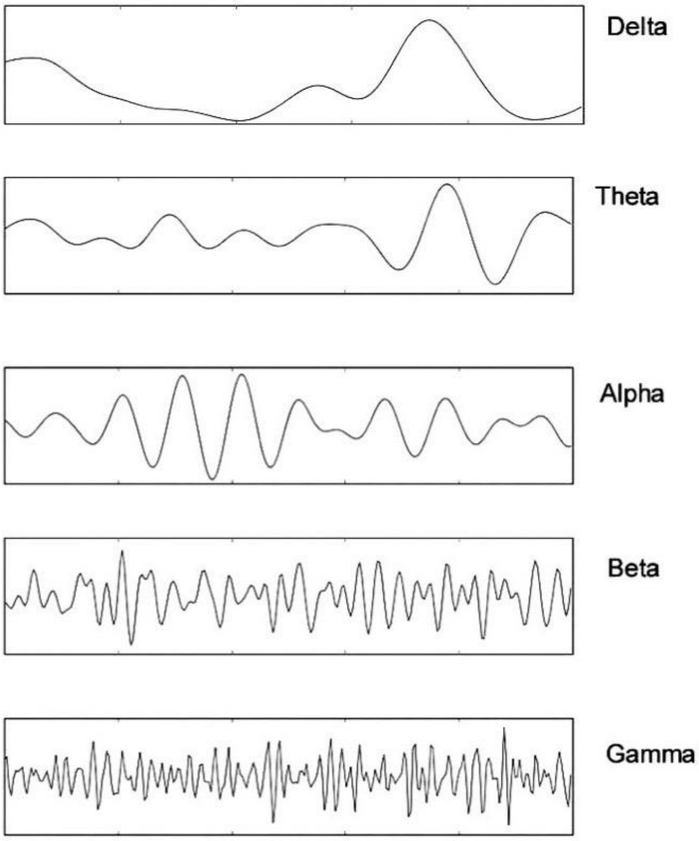


Figure 2-2 Waveforms of different frequency bands for EEG signals[32]

Table 2-1 Introduction to the EEG bands on frequency and cognitive characteristics

Band	Frequency/Hz	Amplitude/μV	Cognitive Characteristics
Delta	0.1-4	10-20	It is often seen in the frontal area of the brain and is common in children and adults during their sleeping. It is less common in normally awake adult brain waveforms. What's more, excessive ventilation, eye opening or calling names could not have a significant effect on the Delta waveform.
Theta	4-8	20-40	The waveforms of the temporal and parietal lobes are more obvious, and generally appear when people are drowsy. It is a manifestation of a state of depression of the central nervous system. It is also associated with working and memory load.
Alpha	8-13	10-100	Alpha is the basic EEG wave in adult brain. Its waveform can be periodically gradually raised and decreased. Alpha is present in all areas of the brain, most obvious in the apical occipital region, Alpha waves are symmetrical on the left and right side of the brain. It can be occurred when people are quiet and with eyes closed. It is generally thought to be related to the brain's preparatory activity.
Beta	13-31	5-30	It is obvious in the forehead, temporal and central area. Beta waveform is more active when people are in concentration or during emotion tension.
Gamma	>32	Unfixed range	It occurs mostly in the frontal and central regions. The increased performance of Gamma waves means the increased excitability of nerve cells. It plays an important role in the reception, transmission, processing, feedback and other advanced functions of received information in the brainstem.

3. Hard to disguise

Different from facial expressions, EEG signals are more objective, hard to disguise, and can reflect the true emotional state of subjects. To identify emotions, researchers need to quantify emotional states. At present, there are two main types of emotion quantification models that are widely used: discrete model and dimensional model. For the discrete model, the emotional space is composed of several emotional states that are discrete, such as happiness, sadness, surprise, fear, anger, and disgust. Those six basic emotional states can make up other emotions[33].

The dimensional model is based on two dimensions: valence-arousal or three dimensions: valence-arousal-dominance[34]. Valence indicates the degree of positivity of emotions, generally from negative to positive. Arousal indicates the intensity of emotions, from weak to strong. Dominance refers to the ability to control emotions, ranging from out of control to in control. The valence-arousal[35] schematic diagram is shown in Figure 2-3.

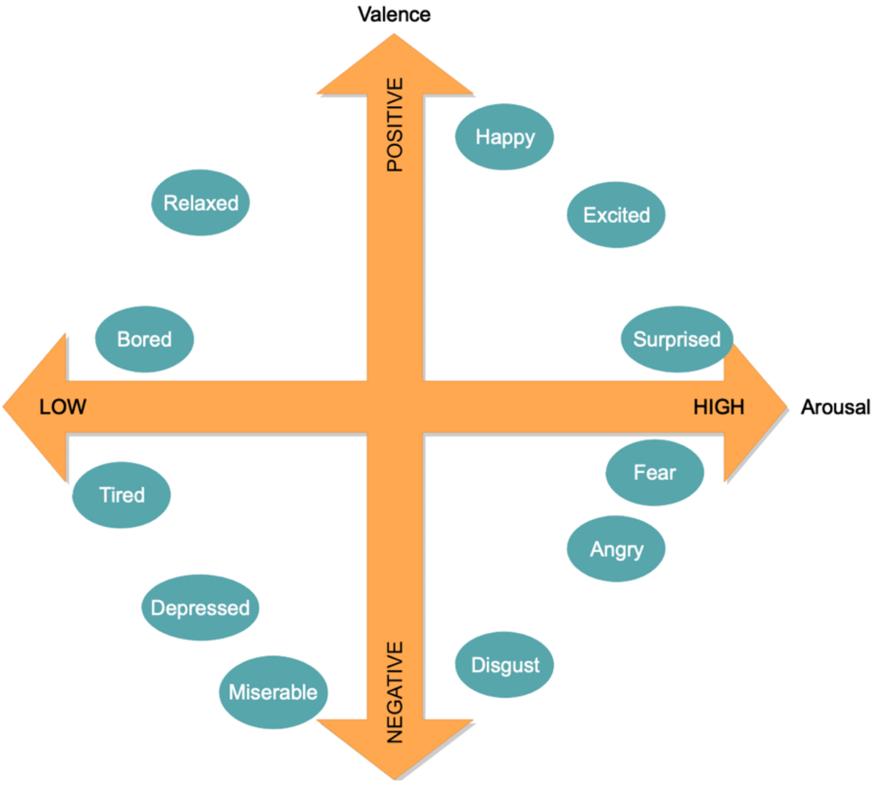


Figure 2-3 The valence-arousal schematic diagram and the corresponding discrete emotions

2.3 DEAP Dataset

DEAP[36] is a database collected by researchers such as Koelstra from several universities. It is usually used to study multi-channel data on human emotional states, which is publicly available. The EEG signals and PPS of 32 subjects are recorded.

For the video stimuli needed in this experiment, the researchers obtained 120 initial video stimuli from the last.FM music website. Of these, 60 music videos are randomly selected, and the remaining half are manually selected. The emotional state corresponding to the selected video corresponds to the two-dimensional emotional space model. In the 120 videos, 40 video stimuli with more pronounced emotional characteristics were finally chosen. Each music video lasts 1 minute.

The EEG signals in the experiment are obtained by 32 active AgCl electrodes with a sampling rate of 512 Hz. The electrodes are placed according to the 10-20 international standard system shown in Figure 2-4. In addition to this, for the first 22 subjects, Sony cameras are used for facial video recording during the experiments. There are total of 32 participants, half of them were males and the others were females, both aging from 19 to 37 years old. The physiological data collected during the experiment are shown in Table 2-2.

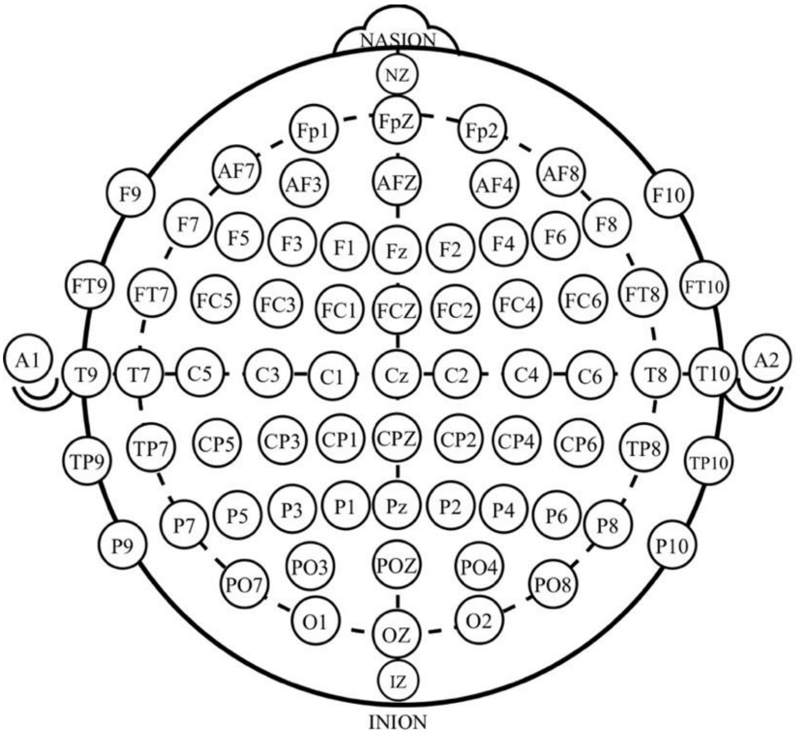


Figure 2-4 Schematic diagram of electrode placement during EEG signal acquisition[37]

Table 2-2 Signal categories and channels collected in DEAP

Signal Category	Channel
EEG	Fpl、FpZ、Fp2、F7、F3、Fz、F4、F8、FT9、FC5、FC1、FC2、FC6、FT10、T7、C3、Cz、C4、T8、CP5、CP1、CP2、CP6、P7、P3、Pz、P4、P8、POZ、O1、OZ、O2
EOG	Horizontal electro-oculogram, hEOG Vertical electro-oculogram, vEOG
EMG	Zygomaticus major electro-myography, zEMG Trapezius electro-myography, tEMG
GSR	Measuring galvanic skin resistance by positioning two electrodes

	on the distal phalanges of the middle and index fingers[36]
BVP	Plethysmograph measures blood volume in the participant’s thumb
Respiration Amplitude	Respiration belt

Each subject is required to collect their physiological data from 40 experiments and record their self-assessment results after each experiment. Before the whole experiment begins, a cross image is displayed on the screen to make the subjects relaxed. A two-minute baseline is recorded during the subject’s emotional calm period and then the experiment begins:

1) The current experiment serial number is displayed on the screen to prompt the subject for the progress of the whole experiment.

2) Display a cross image on the screen for five seconds of baseline recording.

3) Play the 63 seconds music video. The first three seconds are for recording EEG data of each experimental video conversion, and the last sixty seconds are for recording EEG data induced by the video.

4) Subjects finish the self-assessment on valence, arousal, liking, and dominance of each music video based on the self-assessment indexes.

5) After completing 20 experiments, subjects are asked to take a break and prepare for the second half of the experiment.

The data used in this thesis is data_preprocesses_python.zip, which has already been preprocessed, including down sampling and ocular artifacts removal by BBS. The sampling rate of this dataset is 128Hz and the EEG signal frequency band ranges from 4Hz to 45Hz. Each subject’s data consists of 63 seconds’ EEG signals and labels. The data is a matrix of 40*40*8064 and the labels are four kinds: arousal, valence, liking, and dominance.

2.4 EEG Signal Preprocessing

It can be seen that EEG signals are random non-stationary and weak signals from the previous analysis. That’s why there is always noise mixed with acquired signals, such as ECG signal, EMG signal, eye electrical signal, and so on. Therefore, eliminating EEG artifacts to obtain pure EEG signals is the first step for signal analysis. The preprocessing of EEG signals generally contains the following processes:

1. Filtering

EEG acquisition devices usually have their filtering function. The frequency of EEG signals can be controlled before collecting the subjects’ EEG data. Researchers can also use a band-pass filter after acquiring EEG data. The frequency generally ranges from 0.5Hz to 30Hz for emotion analysis. Then the passable frequency range of the band-pass filter can be set from

0.5Hz to 30Hz.

2. Manual denoising

This process is mainly for EEG signals with obvious abnormalities, which are usually caused by limb movements and so on. There will be obvious abnormalities shown in the EEG waveform. It is quite easy to notice this noise and remove them manually.

3. Ocular artifacts removal using BBS

Ocular artifacts[39] refer to the electrical signal generated by eye movements and blinking during EEG signal acquisition. The BBS method could be used to remove ocular artifacts.

BBS[40] means the separation of aliased signals when the theoretical model of signals or the characteristics of the source signal is not accurately known. EEG signals and the artifact part can be separated into independent source signals through the BBS method. Then the artifact signals will be sent directly to zero, we can get relatively pure EEG signals.

The schematic of BBS for EOG artifact removal is shown in Figure 2-5:

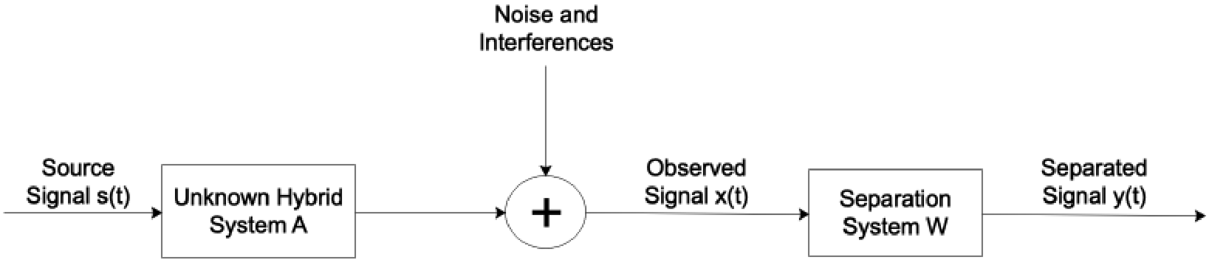


Figure 2-5 The schematic of BBS for artifact removal

The word “blind” in the BBS method means, firstly source signal s(t) is unknown, and secondly how the signals are mixed is unknown.

Independent component analysis (ICA)[41] was gradually developed in the 1990s to solve the BBS problem. It is used in many areas in life such as voice signal processing, communication, and even financial field.

The basic idea of independent component analysis is to get the pure signals extracted from the recorded mixed signals, assuming that the statistical characteristics of the source signals are independent of each other. Its mathematical formula is:

$$x(t) = As(t) + n(t) \tag{2.1}$$

Wherein, A is the transfer matrix of $m * n$ dimension, $x(t)$ is the EEG signals collected in an experiment, $s(t)$ is the statistically independent M $(M \leq N)$ dimension unknown source signal, and $n(t)$ is the observed noise.

The mathematical model of the standard ICA method is a linear noise-free hybrid model. Where the number of EEG source signals N and the number of mixed signals M are equal. For any moment in $x(t) = As(t)$, it has:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_M(t) \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MN} \end{pmatrix} \begin{pmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_N(t) \end{pmatrix} \tag{2.2}$$

The purpose of ICA method is to seek the corresponding linear transformation W, through which the independent source signal $s(f)$ can be separated from acquired signal $x(f)$:

$$u(t) = Wx(t) = WAs(t) \tag{2.3}$$

where $u(t)$ is the estimated vector of s(t), and the transfer matrix A is the inverse matrix of linear transformation W:

$$A = W^{-1} \tag{2.4}$$

The basic steps of the ICA based ocular artifact removal method are:

1) Effectively separate the source signals whose statistical characteristics are independent of the acquired EEG signal $x(t)$ by linear transformation W. That is to achieve the effective separation of the ocular artifact source signal and the actual EEG signal.

2) Remove the separated EEG ocular artifact source signals, which are being set to amplitude zero, and obtain the estimated source signal $u'(t)$. $u'(t)$ is the actual EEG source signal.

3) Calculate $x'(t) = w^{-1}u'(t)$ to obtain the pure EEG signals.

ICA method can accurately identify and effectively remove ocular artifact interferences. The example schematic diagram after ICA processing is shown in Figure 2-6 and Figure 2-7.

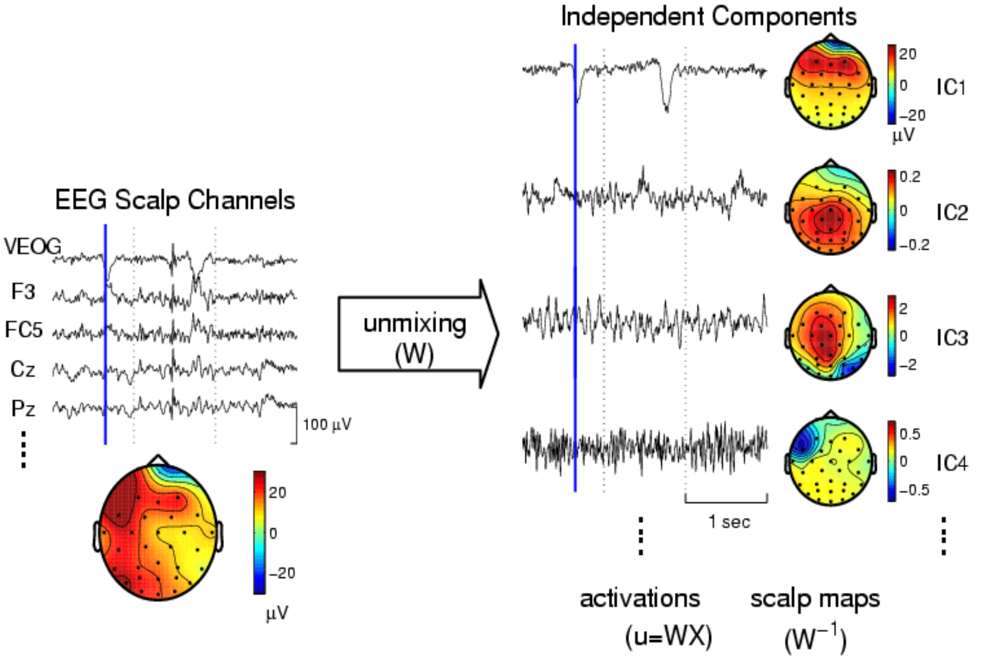


Figure 2-6 Schematic diagram of the independent components of EEG signals using ICA[42]

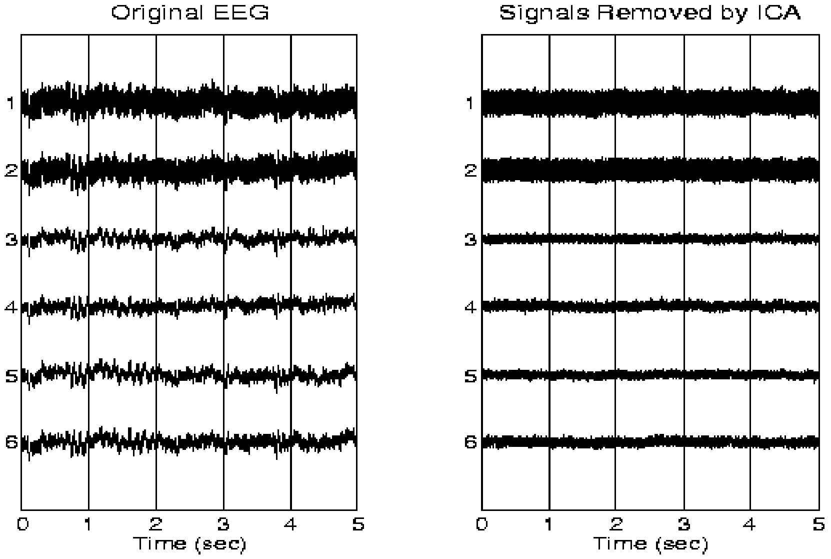


Figure 2-7 EEG waveform diagram after ICA processing[43]

However, there are still some main problems with such an algorithm[44]:

1) The identification of ocular artifacts requires a lot of experience accumulation and manual intervention, which is subjective. The ocular artifact identification is not highly automatic and it's time-consuming.

2) High accuracy may cause increased complexity. The algorithm requires a sufficient number of channels when collecting EEG signals, but a large number of channels will greatly increase the complexity of the algorithm.

3) It is easy to cause the loss of power spectrum of EEG signals.

2.5 Feature Extraction Using Wavelet Transform

The wavelet transform[45] is a time-frequency analysis method with variable resolution. It is the inheritance and development of the Fourier transform. The wavelet transform decomposes a signal into a series of parent wavelet functions or a superposition of wavelet basis functions. When analyzing low-frequency signals, the time window of the wavelet transform is large, which is characterized by high-frequency resolution and low time resolution. When analyzing high-frequency signals, the time window of the wavelet transform is small, which is characterized by low-frequency resolution and high time resolution. Wavelet transform has a wide range of applications in signal processing, image processing, speech recognition, as well as in data compression, error diagnosis, and quantum physics.

In general, the definition of the wavelet transform is as follows:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right) \tag{2.5}$$

In this formula, a is the scaling parameter and b is the translation parameter.

In signal processing, the continuous wavelet transform needs to calculate all the parameters at all possible scales, so it will generate a large amount of redundant information and occupy computer resources. In order to solve this problem, a mutually orthogonal basis function is considered. The scaling parameter "a" and the translation parameter "b" in the continuous wavelet transform definition formula is discretized to obtain a new discrete wavelet function. The definition of the discrete wavelet transform is as follows[46]:

$$DWT(j,k) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{2j}} \psi\left(\frac{t-2^j k}{2^j}\right) dt \tag{2.6}$$

In the formula above, $2^j k$ is the time parameter, 2^j is the scale parameter, and $\psi(t)$ is the parent wavelet function. The four layers wavelet decomposing schematic diagram of EEG can be seen in Figure 2-8.

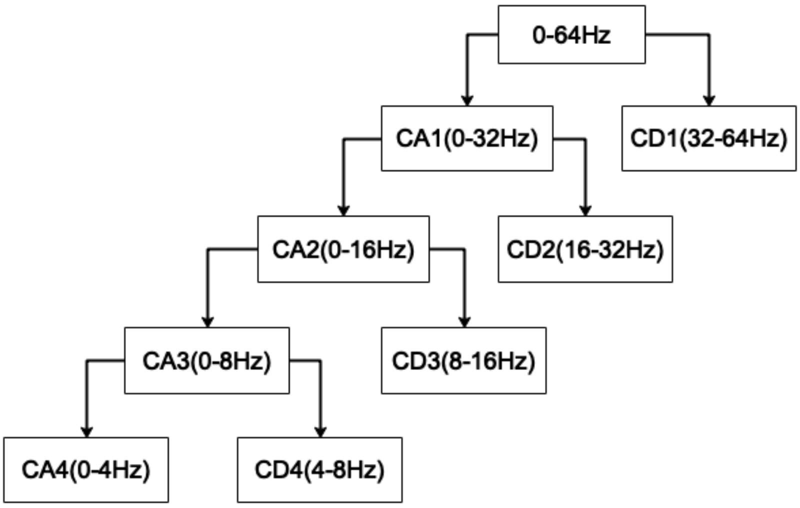


Figure 2-8 Four layers wavelet decomposing schematic diagram of EEG signals

The EEG emotion recognition database used in this experiment is the data after preprocessing (including down sampling, ocular artifact removal by BBS). The sampling frequency of EEG signals is 128Hz, and the signal frequency ranges from 3Hz to 45Hz. We can know that the detectable signal band range is 0-64Hz according to Nyquist's sampling theorem. The result of four layers of wavelet decomposition of EEG signals using discrete wavelet transforms is shown in Figure 2-9.

In Figure 2-9, the signal in the first row is the original EEG of a subject in a certain channel, the signal in the second row is the approximated component in level 4, the signal in the third row is the detained component in level 4. The signal in the fourth row is the detailed component in level 3, the signal in the fifth row is the detained component in level 2 and the last row is the detailed component in level 1. Theta band frequency ranges from 4Hz to 8Hz, located on the detailed component CD4. Alpha band frequency ranges from 8Hz to 13Hz, located on the detailed component CD3. Beta band frequency ranges from 14Hz to 30Hz, located on the detailed component CD2 and Gamma band frequency ranges from 30Hz to 45Hz, located in the detailed component CD1.

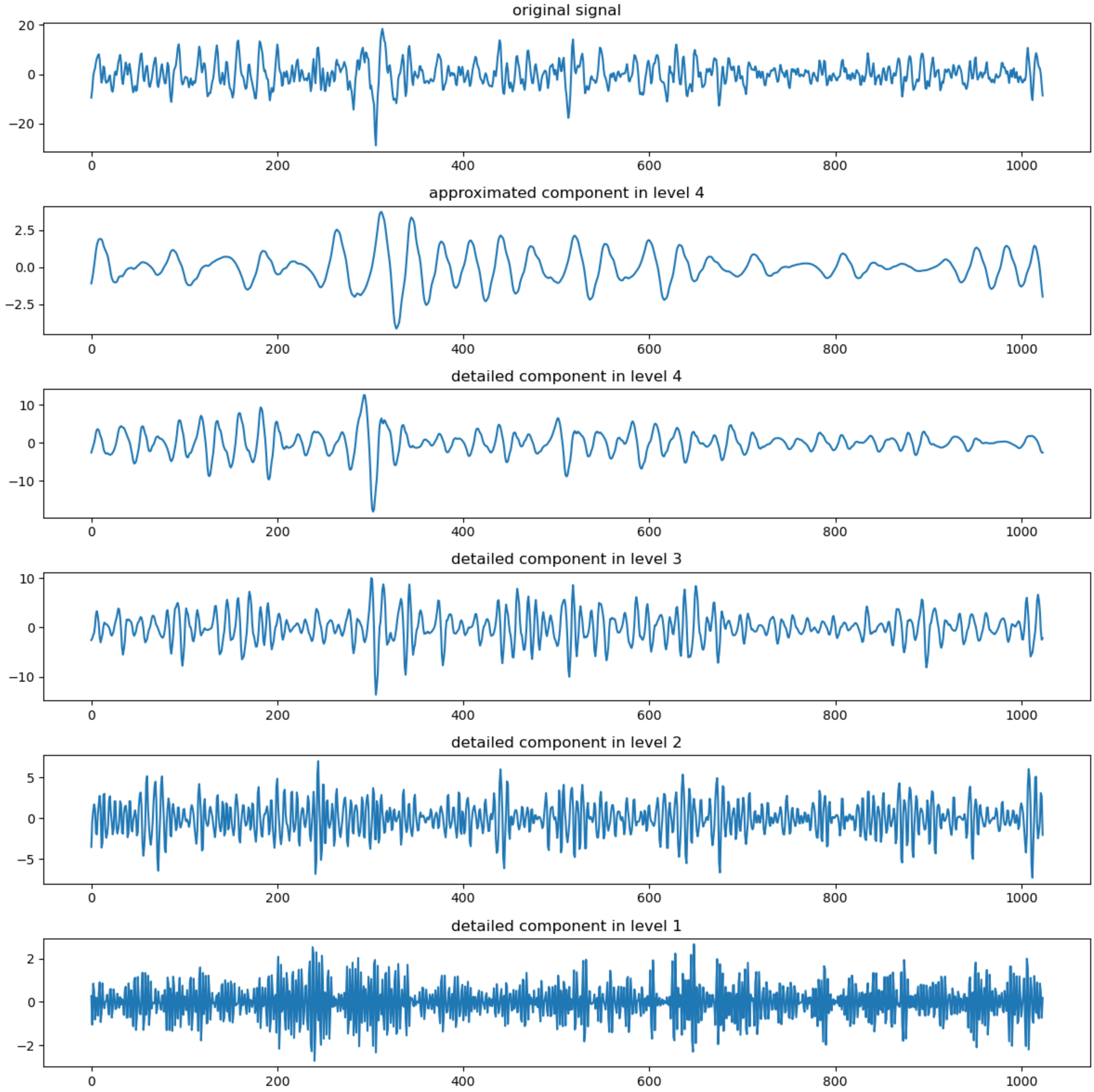


Figure 2-9 Four layers decomposing using wavelet transform

2.6 Proposed SVM-PCA for Emotion Recognition

The SVM is a classification model of supervised learning. It is proposed by Vapnik based on statistical learning theory[47]. The basic model of a SVM defines a linear classifier with the largest distance in the feature spacs[48]. It can also be changed into a nonlinear classifier. If the training samples are linearly inseparable in low-dimensional space, the SVM kernel can be used to map them to a high-dimensional space, implementing the transformation of the samples from linear inseparable to non-linear separable, which is shown in Figure 2-10.

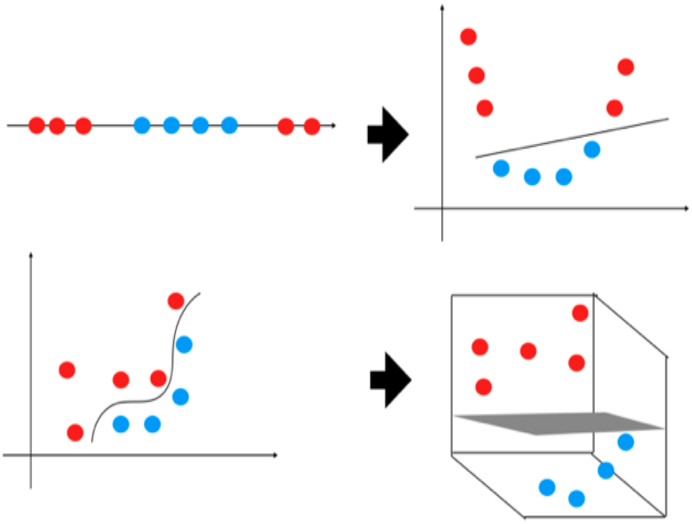


Figure 2-10 Data mapping from low dimension to high dimension[49]

Taking the two-dimensional space as an example, the algorithm principle of the SVM is introduced in Figure 2-11.

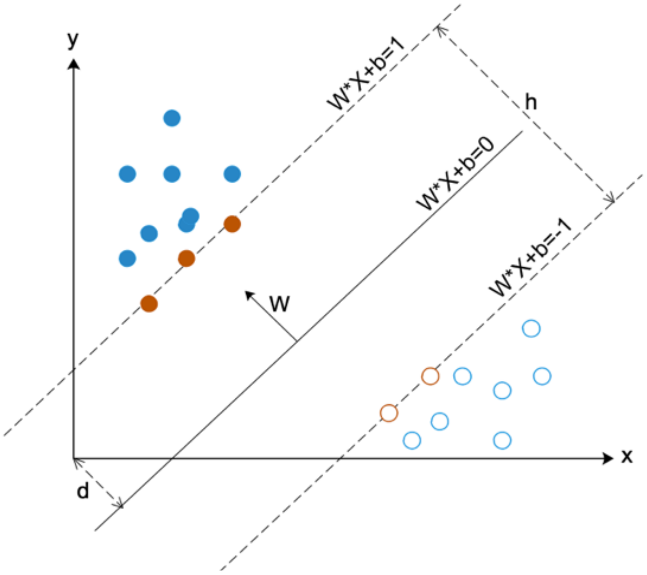


Figure 2-11 SVM model of two-dimensional plane

In Figure 2-11, $w \cdot x + b = 0$ is the separating hyperplane. For a linearly separable dataset, there is an infinite number of such hyperplanes (i.e. perceptron). But the separating hyperplane with the largest geometric intervals is unique. SVM algorithm is to find the hyperplane that can correctly divide the training dataset and have the largest geometric intervals.

Suppose there is a training dataset on a feature space:

$$\begin{gathered} T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_N, y_N)\}, \\ \boldsymbol{x}_i \in R^n, y_i \in \{-1, +1\}, i = 1,2,\dots,N \end{gathered} \tag{2.7}$$

$\boldsymbol{x}_i$ is the ith eigenvector and y_i is the class marker. When the training data is above the hyperplane, its corresponding result is defined as +1, and when the training data is below the hyperplane, its corresponding result is defined as -1.

For a given dataset T and the hyperplane $w \cdot x + b = 0$, the definition of the geometric interval of the hyperplane concerning the samples is

$$\gamma_i = y_i \left(\frac{\boldsymbol{w}}{||\boldsymbol{w}||} \cdot \boldsymbol{x}_i + \frac{b}{||\boldsymbol{w}||} \right) \tag{2.8}$$

The minimum geometric interval on the hyperplane for all sample points is:

$$\gamma = \min_{i=1,2,\dots,N} \gamma_i \tag{2.9}$$

γ can be considered the distance from the support vector to the hyperplane. According to the above definitions, finding the hyperplane with the largest geometric intervals can be transformed into solving the optimization problem below:

$$\begin{cases} \max\limits_{w,b} \gamma \\ s.t. \quad y_i \left(\frac{\boldsymbol{w}}{||\boldsymbol{w}||} \cdot \boldsymbol{x}_i + \frac{b}{||\boldsymbol{w}||} \right) \geq \gamma, \quad i = 1,2,\dots,N \end{cases} \tag{2.10}$$

For the sake of conciseness of the formula, let $\boldsymbol{w} = \boldsymbol{w}/||\boldsymbol{w}||\gamma$ and $\boldsymbol{b} = \boldsymbol{b}/||\boldsymbol{w}||\gamma$, then we can have $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, i = 1,2,\dots,N$. Maximizing γ is equivalent to minimizing $(1/2) \cdot ||\mathbf{w}||^2$. Therefore, the solution of the hyperplane for SVM model can be transformed as the following constraint optimization problem:

$$\begin{cases} \min\limits_{w,b} \frac{1}{2} ||\boldsymbol{w}||^2 \\ s.t. \quad y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, \quad i = 1,2,\dots,N \end{cases} \tag{2.11}$$

With the introduction of the Lagrange multiplier, the above formula becomes:

$$\begin{cases} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{||\boldsymbol{w}||^2}{2} - \sum\limits_{i=1}^{N} \alpha_i [y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - 1] \\ s.t. \quad \alpha_i \geq 0, \quad i = 1,2,\dots,N \end{cases} \tag{2.12}$$

α_i is Lagrange multiplier. Using the duality of Lagrange function, we can turn an optimization problem into a problem for solving minimum:

$$\begin{cases} \min\limits_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i \cdot \boldsymbol{x}_j) - \sum_{i=1}^{N}\alpha_i \\ s.t. \quad \sum_{i=1}^{N}\alpha_i y_i = 0, \ \ \alpha_i \geq 0, \ \ i = 1,2,\cdots,N \end{cases} \tag{2.13}$$

When there is a classification-optimal solution to the training set, the corresponding function is:

$$f(\boldsymbol{x}) = sgn(\boldsymbol{w}^* \cdot \boldsymbol{x} + b^*) = sgn\left(\sum_{i=1}^{N}\alpha_i^* y_i x_i^* x + b^*\right) \tag{2.14}$$

In the above formula, we have:

$$\boldsymbol{w}^* = \sum_{i=1}^{N}\alpha_i^* y_i \boldsymbol{x}_i \tag{2.15}$$

$$b^* = y_j - \sum_{i=1}^{N}\alpha_i^* y_i(\boldsymbol{x}_i \cdot \boldsymbol{x}_j) \tag{2.16}$$

$$sgn(t) = \begin{cases} 1, & t > 0 \\ 0, & t = 0 \\ -1, & t < 0 \end{cases} \tag{2.17}$$

α_i^* is the support vector for the classification label, and b* is the threshold for the classification. In specific application scenarios, most of the training samples used in SVM are linear inseparable. To solve this problem, we need to use kernel functions. A commonly used kernel function is the Gaussian kernel function. It has the advantages of few parameters, high flexibility, and strong immunity to interference. The expression of Gaussian kernel function is as follows:

$$K(x,z) = exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right) \tag{2.18}$$

In this case, the classification decision function is:

$$f(x) = sgn\left(\sum_{i=1}^{N}\alpha_i^* y_i exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right) + b^*\right) \tag{2.19}$$

PCA is a kind of data analysis technology, the main idea is to project high-dimensional data into a lower dimensional space and extract the main features. PCA has a wide range of applications and is often in conjunction with other methods such as classification, clustering, and data processing. It can efficiently identify the main features and reduce the dimensionality of the original data. The goal of PCA is to find x $(x < n)$ new variables, use these x variables to reflect the main features, and achieve compression of the original data matrix size. These x

new variables are the "principal components" that reflect the characteristics of the original n variables, and the x variables are uncorrelated with each other. Using PCA, a part of the principal components is discarded and only the first few principal components with large variance are chosen to represent the original data, which can reduce the calculation.

If you want PCA to have better performance, a sufficient amount of data is needed for training. What's more, when using PCA, we must also pay attention to data normalization. Different base units will cause different proportions of variables when calculating eigenvalues, and it may affect calculation results.

Considering that PCA could use fewer dimensions while retain most of the data features and its excellent running speed, we combine PCA with SVM. The results of accuracy on the four labels using both SVM and SVM-PCA for emotion recognition are shown in Figure 2-12.

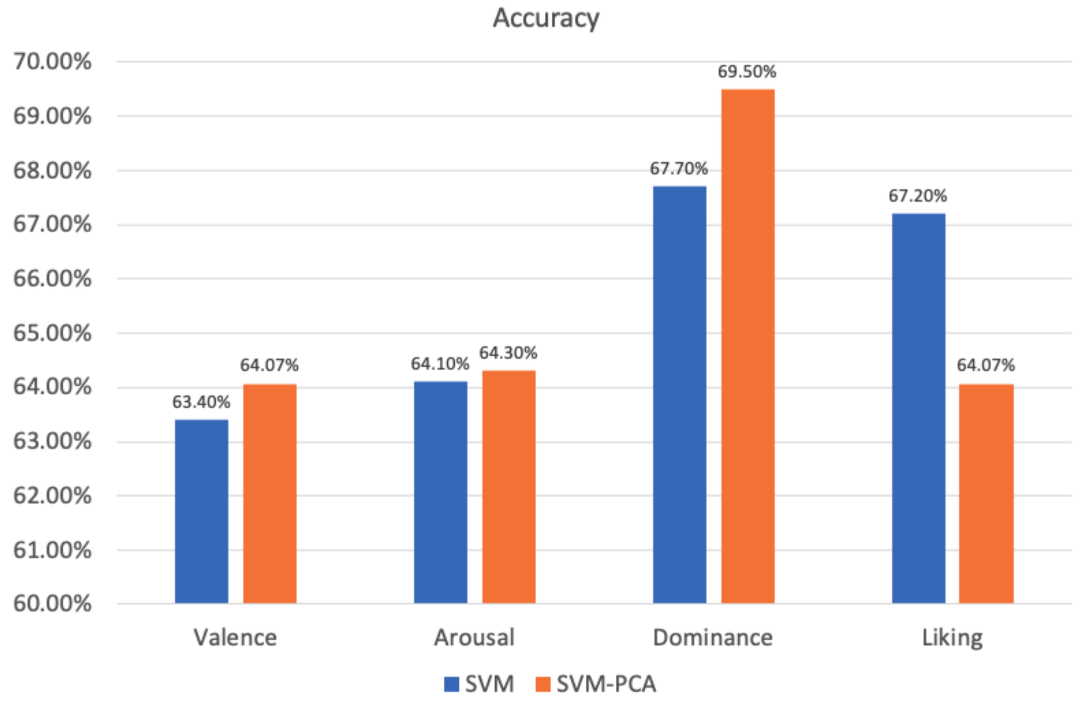


Figure 2-12 The accuracy of four labels using SVM and SVM-PCA for emotion recognition

2.7 Chapter Summary

The generating mechanism, features, frequency band distribution, and emotion classification theory of EEG signals are all covered in this chapter. It also introduces the basic content of the DEAP dataset used in this thesis, including the collection method of EEG signals and the collection of facial expressions. The preprocessing procedures for EEG signals are then explained, including filtering, manual denoising, and the use of the ICA approach to remove ocular artifacts. The process of decomposition and feature extraction using wavelet transforms is then introduced. Finally, using EEG data from the DEAP dataset, the SVM network is trained and tested, with the accuracy of the test results on arousal and valence reaching roughly 64%.

3 Proposed Deep-CNN for Emotion Recognition Based on Facial Images

3.1 Basic Steps for Emotion Recognition Based on Facial Images

In facial emotional recognition part, the video recording data from the first twenty-two subjects contained in DEAP dataset is analyzed. Firstly, the facial expression images are captured from the videos. Every 100 frames of a one-minute video, one image is captured (every two seconds). Secondly, the Adaboost algorithm is used to detect faces and segment images. The segmented facial images are converted into a matrix of 48*48. This image dataset is split into two parts: training and testing, and it is fed into a deep-CNN for training and determining emotion recognition accuracy. The flowchart is shown in Figure 3-1.

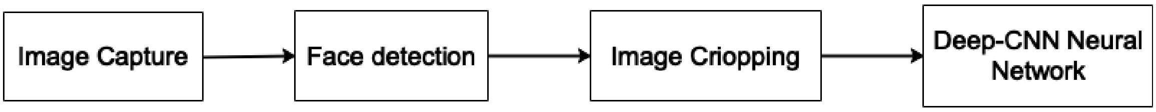


Figure 3-1 The flowchart for emotion recognition based on facial expressions

3.2 Face Detection Using Adaboost Algorithm

In DEAP dataset, the video recording during the experimental process is only performed on the first 22 subjects. Each of the subjects is asked to watch 40 music videos, so there will be 40 one-minute videos for each of them. First of all, a total of 880 videos are converted into images. It is captured every 100 frames (every two seconds). That means, for each one-minute video, there are 30 corresponding facial images.

Face detection refers to the positioning of the face in the image. If there is a face in the image, then obtain the position, the size, and the posture information of the face. Face detection is the basis for further analysis and understanding of facial expressions.

There are many face detection algorithms, but the most used one is the AdaBoost algorithm proposed by Paul Viola in 2001[50]. The detection accuracy, speed, robustness, and other performance of the algorithm have reached the requirements, making face detection more practical. The core idea of this algorithm is to use the integral image method to calculate Haar-like[51] features of different scales and directions, then use the Adaboost algorithm to construct a classifier. In real-time face detection, the cascading structure is commonly used. The strong classifier after the iteration of the weak classifier is constructed into a cascading classifier, which can quickly exclude the non-face area and implement face detection.

In other words, the Adaboost face detection method can be divided into the following three steps:

1) The Haar-like features are used to represent faces. The integral image method is for the rapid calculation of eigenvalues.

2) The Adaboost will first use some weak classifiers to select features that can represent human faces. The different weak classifiers are then combined to construct a strong classifier.

3) Several strong classifiers consisting of weak classifiers are connected to obtain a new polar classifier. Such a structure can effectively increase the face detection accuracy.

Haar-like features were originally proposed by Papageorgiou et al.[52] when they are using Haar wavelet transform to extract features from human faces. Haar-like features usually consist of 2 to 4 rectangles to detect boundary, line, or diagonal features, respectively, as shown in Figure 3-2.

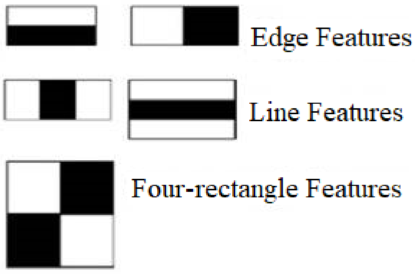


Figure 3-2 Haar-like features[53]

We use three types of features here. Detecting edge features uses the first two features of two rectangles shown in the figure above. The eigenvalue is the difference between the sum of pixel values in the white rectangle and the sum of pixel values in the black rectangle. Detecting line features uses the third and fourth features, each containing three rectangles. The feature is defined as the difference between the sum of pixel values in two white rectangles and pixel values in the black rectangle. Detecting diagonal features utilizes the fifth feature, which includes four rectangular parts. The feature is defined as the difference between the sum of pixel values in two white rectangles and the sum of pixel values in two black rectangles. The number of Haar-like features is quite large because one Haar-like feature represents a scale and a location of eigenvalues. We could use the integral image method to do the fast calculations on such a large number of features.

The Adaboost algorithm[54] can be used for feature selection, based on the assumption: there are a small number of features in many Haar-like features, and combining them to get an efficient classifier. For a weak classifier y_m, suppose the total training sample number in sample set S is N, and the maximum number of training rounds is M.

The specific steps of the Adaboost algorithm are described below:

1) Initialize samples $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ in sample set S, and each with a weight of $1/N$.

2) Go through M loops and initialize the first iteration m $=$ 1.

Train the weak classifier y_m and the weight error minimum function is:

$$\varepsilon_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(X_n) \neq t_n) \tag{3.1}$$

Calculate the error rate of the weak classifier:

$$\alpha_m = ln\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right) \tag{3.2}$$

Select the appropriate threshold function to minimize the error ε_m.

Update the weights of the weak classifier:

$$w_{m+1,i} = \frac{w_{mi}}{Z_m} exp(-\alpha_m t_i y_m(x_i)), i = 1,2, \ldots, N \tag{3.3}$$

where,

$$Z_m = \sum_{i=1}^{N} w_{mi} exp(-\alpha_m t_i y_m(x_i)) \tag{3.4}$$

$Z_m$is the normalized factor that gives the sum of the weights of all weak classifications to 1.

3) After an M-iteration loop, M weak classifiers and different weak classifier weights are obtained. The finally assigned weights are combined to get a strong classifier, and the functional expression of the strong classifier is:

$$Y_M(x) = sgn(\sum_{m=1}^{M} \alpha_m y_m(x)) \tag{3.5}$$

From the above content, it is clear that the sum of the weights of all weak classifiers is 1. If one of the weak classifiers is assigned the wrong weight, it will inevitably cause an imbalance in the weights of the other classifiers. The emergence of face detection based on the Adaboost algorithm fundamentally solves the slow detection problem. It can locate the face in the image extremely quickly and the algorithm follows the steps:

1) Using the concept and calculation method of the integral image, the eigenvalues of Haar-like can be calculated quickly.

2) Feature selection is carried out according to the Haar-like features, then a weak classifier is formed. Finally, the weak classifiers are assigned with different weights to form the final strong classifier.

3) Cascade[55] is used to iterate according to each strong classification, which improves the detection speed.

The process of face detection using the Adaboost Cascade classifier is shown in Figure 3-3. An example of Subject 1 of face detection result using Adaboost algorithm is shown in Figure 3-4 and the image cropping result is shown in Figure 3-5.

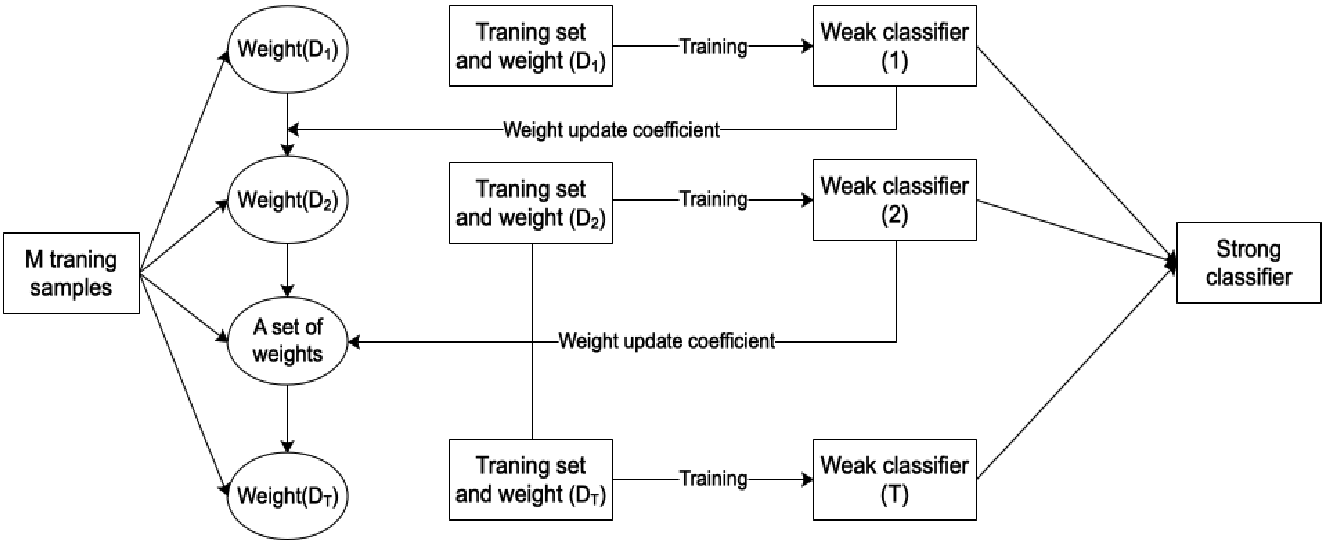


Figure 3-3 The diagram of Adaboost Cascade classifier

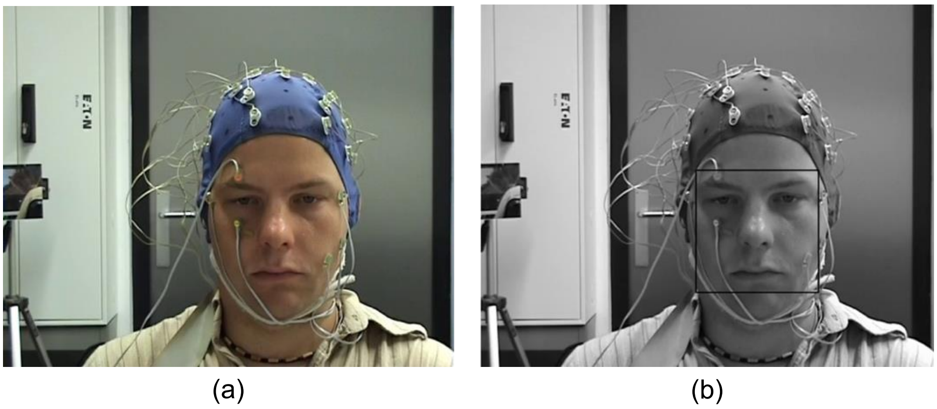


Figure 3-4 (a) An example of Subject 1 captured image from the recorded video
(b) The face detection result using Adaboost algorithm for captured image

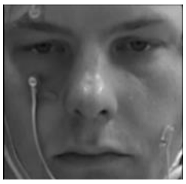


Figure 3-5 Face cropping result after face detection

3.3 Facial Image Preprocessing

Colorful images have a large amount of irrelevant information and more interference information and are not suitable for facial expression recognition. For the facial expression recognition system, grayscale images can fully meet the requirements. Therefore, grayscale images[56] are used for emotion recognition in this thesis.

The segmented facial images are converted from RGB composite channel image to 8-bit grayscale image. The converted image is divided into 0 to 255 for a total of 256 gray levels. 0 represents the darkest and 255 represents the brightest. The conversion formula is:

$$Gray = 0.229 \times R + 0.587 \times G + 0.114 \times B \tag{3.6}$$

In a real system, in order to speed up the conversion, the formula is improved to:

$$Gray = (77 \times R + 151 \times G + 28 \times B) \gg 8s \tag{3.7}$$

The size of the face area after face detection and image segmentation is different, in order to meet the needs of emotion recognition, the face area images should be uniformly normalized to the same size. If the normalized face image size is too large, on the one hand, the emotion feature extraction time will increase, on the other hand, the extracted feature vector dimension will be too high. It may also lead to that the algorithm' efficiency and accuracy rate can not achieve satisfactory results. If the normalized face image size is small, the over-compressed images will lose a lot of facial expression information, resulting in a decrease in accuracy.

When the eye position is certain, the facial expression area can also be found. Using the center of the left and right edges of the face and the position of the eyes as baseline, the image can be cropped to filter out useless information outside the area of the face. Each image size is 48*48 pixels.

There are many algorithms for image size adjustment, such as nearest neighbor interpolation[57], bilinear interpolation, etc. Considering the images after bilinear interpolation has the high image quality and good continuity, bilinear interpolation is selected for dimensional normalization.

Bilinear interpolation[58], also known as first-order interpolation, is an application for linear interpolation which is extended to two dimensions. Four adjacent points around the point (x, y) form a unit square. The coordinates of these four vertices are $f(0,0)$, $f(0,1)$, $f(1,0)$, $f(1,1)$. To obtain the pixel value $f(x, y)$ of any point (x, y) in the square, the following steps are:

By linear interpolation of the two points at the upper end, it has:

$$f(x, 0) = f(0,0) + x[f(1,0) - f(0,0)] \tag{3.8}$$

By linear interpolation of the two vertices at the lower end, it has:

$$f(x, 1) = f(0,1) + x[f(1,1) - f(0,1)] \tag{3.9}$$

By linear interpolation in the vertical direction, it has:

$$f(x,y) = f(x,y) + y[f(x,1) - f(x,0)] \tag{3.10}$$

Combine three formulas above, we can get:

$$f(x,y) = [f(1,0) - f(0,0)]x + [f(0,1) - f(0,0)]y + \\ [f(1,1) + f(0,0) - f(0,1)]xy + f(0,0) \tag{3.10}$$

After the facial expression images are normalized by scale, they are consistent in geometric space. However, the grayscale image has a large grayscale distribution range under different conditions. The large gray scale span affects the richness of facial expression. In order to remove this undesirable factor, the expression images need to be gray equalized. The histogram equalization method is used in this thesis.

Histogram equalization is a method of enhancing image contrast, the main idea of which is to change the histogram distribution of an image into an approximate uniform distribution, thereby enhancing the contrast of the image. Although histogram equalization is only a basic method in digital image processing, it is very powerful and is a very classic algorithm.

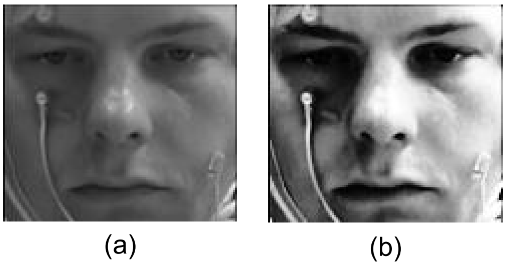
Assume the facial expression image has L levels grayscale (0,1, 2, …, L-1). N is the number of pixels in the image, and n_i is the number of pixels in the ith grayscale level. The probability of pixels in the image with grayscale i is:

$$P(i) = \frac{n_i}{N} \tag{3.11}$$

Once the grayscale histogram is obtained, the grayscale equalization[59] can be carried out. The conversion formula for histogram equalization is:

$$b = T(a) = (L-1)\sum_{i=0}^{k} P(i) = (L-1)\sum_{i=0}^{k} \frac{n_i}{N} \tag{3.12}$$

where a is the gray level of the pixel before conversion and b is the gray level after conversion. The following Figure 3-6 shows the facial image after grayscale equalization.



(a) (b)

Figure 3-6 (a) Cropped facial image before grayscale equalization

(b) Cropped facial image after grayscale equalization

3.4 Proposed Deep-CNN for Emotion Recognition

In 1943, psychologist McCulloch and mathematician Pitts proposed the mathematical model of neurons, which have the function of logical operations, and it is called the MP model[60]. In 1958, Rosenblatt proposed perceptron models and algorithms[61]. In 1968, Rumelhart and Hinton et al.[62] proposed backpropagation algorithms to solve the complex computational quantity problem of multilayer perceptron, which in turn promoted further research on multiplayer perceptron. Then LeCun et al. proposed the LeNet-5 model, which is what we currently known as CNN. However, the amount of computation required for this kind of neural network was very large, and computer calculation ability at that time can not meet the requirement. Until 2006, Hinton proposed a multilayer neural network based on its predecessors. This network has the ability to learn features and can reduce the complexity of its calculations by initializing the weights.

3.4.1 Multilayer Neural Network Model

For supervised learning problems with samples that have been labeled (x_i, y_i), the neural network can define a complex and nonlinear hypothesis[63]: $h_{w,b}(x)$, where w represents the weights of the neural network and b represents the bias. The parameters w and b can be adjusted to be suitable for different dataset. The structure of a single neuron is shown in Figure 3-7:

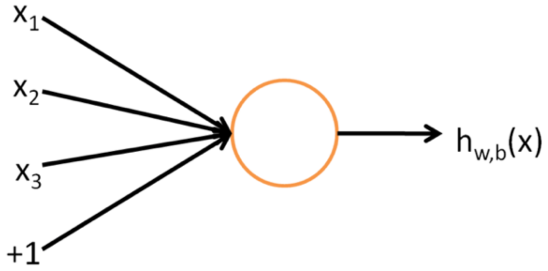


Figure 3-7 The structure of a single neuron

The neuron here is a computational unit with input of x_1, x_2, x_3 and an offset term. The output is:

$$h_{w,b}(x) = f(W^T x) = f\left(\sum_{i=1}^{3} W_i x_i + b\right) \tag{3.13}$$

f is called as activation function. We can select, for example, the Sigmoid function as the activation function, the expression of which is as follows:

$$f(z) = \frac{1}{1 + exp(-z)} \tag{3.14}$$

It is also possible to select, for example, the hyperbolic tangent function as activation function, which is formulated as:

$$f(z) = tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{3.15}$$

A neural network is a structure that connects many simple neurons together. In this way, the neuron's output in the previous layer can be the neuron's input in the next layer. A schematic diagram of a small neural network is in Figure 3-8:

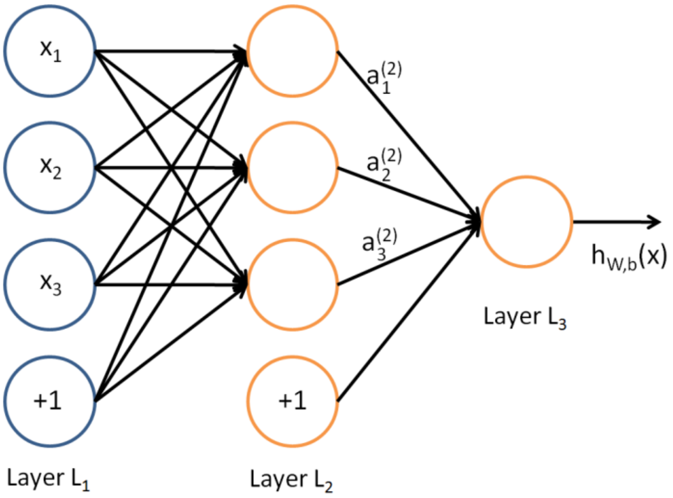


Figure 3-8 The schematic diagram of a small neural network

In the preceding figure, circles are used to represent the nodes of the network. The circle labeled 1 represents the bias term, the leftmost node layer is called the input layer and the rightmost node layer is called the output layer. The middle node layer is called the hidden layer because its node values cannot be observed in the training set. There are three input units, three hidden units, and one output unit in this small neural network example.

The number of neural network layers is represented by n_l. In this example, $n_l = 3$. The lth layer in the neural network is represented by L_l, so L_1 represented the input layer and L_{n_l} represents the output layer. The parameters of neural network are expresses here as:

$$(W, b) = \left(W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\right) \tag{3.16}$$

$W_{ij}^{(l)}$ is the weight connecting the jth unit in the lth layer and the ith unit in the $(l + 1)$th layer. $b_i^{(l)}$ is the bias of the ith unit in the $(l + 1)$th layer. In the example of small neural network:

$$W^{(1)} \in R^{3\times3},$$
$$W^{(2)} \in R^{1\times3} \tag{3.17}$$

Use $a_i^{(l)}$ to represent the output value after activation of jth unit in the lth layer. Therefore, the calculation formula of the neural network can be written as:

$$a_1^{(2)} = f\left(W_{11}^{(1)}x_1 + W_{12}^{(1)}x_2 + W_{13}^{(1)}x_3 + b_1^{(1)}\right) \tag{3.18}$$

$$a_2^{(2)} = f\left(W_{21}^{(1)}x_1 + W_{22}^{(1)}x_2 + W_{23}^{(1)}x_3 + b_2^{(1)}\right) \tag{3.19}$$

$$a_3^{(2)} = f\left(W_{31}^{(1)}x_1 + W_{32}^{(1)}x_2 + W_{33}^{(1)}x_3 + b_3^{(1)}\right) \tag{3.20}$$

$$h_{W,b}(x) = a_1^{(3)} = f\left(W_{11}^{(2)}a_1^{(2)} + W_{12}^{(2)}a_2^{(2)} + W_{13}^{(2)}a_3^{(2)} + b_1^{(2)}\right) \tag{3.21}$$

To simplify the above formula, we use:

$$z_i^{(l+1)} = \sum_{j=1}^{n} W_{ij}^{(l)}x_j + b_i^{(l)} \tag{3.22}$$

If the output value $a^{(l)}$ of the lth layer is known, the output value $a^{(l+1)}$ of the $(l+1)$th layer can be calculated:

$$z^{(l+1)} = W^{(l)}a^{(l)} + b^{(l)} \tag{3.23}$$

$$a^{(l+1)} = f\left(z^{(l+1)}\right) \tag{3.24}$$

This step is called forward propagation. Parameters can be combined into matrix form to perform the calculations in a neural network using fast linear algebra methods.

3.4.2 Backpropagation

For a single training sample, we can define the cost function on this sample as[64]:

$$J(W,b;x,y) = \frac{1}{2}\left\|h_{W,b}(x) - y\right\|^2 \tag{3.25}$$

For a training set containing m samples, the cost function can be defined as:

$$J(W,b) = \left[\frac{1}{m}\sum_{i=1}^{m} J\left(W,b;x^{(i)},y^{(i)}\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2 \tag{3.26}$$

$$= \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{w,b}\left(x^{(i)}\right) - y^{(i)}\right\|^2\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2 \tag{3.27}$$

In the above formula, the first item is a mean squared error term, and the second item is a regularization term, also known as a weight decay term. The weight decay term reduces the wrights and help prevent overfitting. The weight decay parameter λ controls the weight relationship between the two terms. In this formula, $J(W,b;x,y)$ only represents the squared error cost function of a single sample, while $J(W,b)$ is the total cost function. Our goal is to minimize the cost function $J(W,b)$. Initialize each parameter $W_{ij}^{(l)}$ and $b_i^{(l)}$ to a random number close to zero, such as 0.01.

The optimization algorithm of gradient descent is then used to train the neural network. Since $J(W,b)$ is a non-convex function, gradient descent algorithm tends to fall into a local optimal solution rather than the global optimal solution. Therefore, in practical applications, it is necessary to randomly initialize parameters instead of directly setting them to zero. The gradient descent update formula for parameter $W_{ij}^{(l)}$ and $b_i^{(l)}$ is as follows:

$$W_{ij}^{(l)} = W_{ij}^{(l)} - \alpha \frac{\partial}{\partial W_{ij}^{(l)}} J(W,b) \tag{3.28}$$

$$b_i^{(l)} = b_i^{(l)} - \alpha \frac{\partial}{\partial b_i^{(l)}} J(W,b) \tag{3.29}$$

α represents the learning rate. To calculate the partial derivative in the above formula, the backpropagation algorithm is introduced. the specific steps of back propagation algorithm are as follows:

For a given training sample (x,y), a forward propagation process is first performed for calculating the output value after activation from Layer L_2 to Layer L_{n_l}.

For each output unit i in the output layer L_{n_l}, there is:

$$\delta_i^{(n_l)} = \frac{\partial}{\partial z_i^{(n_l)}} \frac{1}{2} \left\| y - h_{W,b}(x) \right\|^2 = -\left(y_i - a_i^{(n_l)}\right) \cdot f'^{\left(z_i^{(n_l)}\right)} \tag{3.30}$$

For each node i in layer l, there is:

$$\delta_i^{(l)} = \left(\sum_{j=1}^{s_{l+1}} W_{ji}^{(l)} \delta_j^{(l+1)} \right) f'^{\left(z_i^{(n_l)}\right)}, l = 2,3,\cdots,n_l - 1 \tag{3.31}$$

Calculate the partial derivative:

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W,b;x,y) = a_j^{(l)} \delta_i^{(l+1)} \tag{3.32}$$

$$\frac{\partial}{\partial b_i^{(l)}} J(W,b;x,y) = \delta_i^{(l+1)} \tag{3.33}$$

The partial derivative of the global cost function is calculated as follows:

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W,b) = \left[\frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial W_{ij}^{(l)}} J\left(W,b;x^{(i)},y^{(i)}\right) \right] + \lambda W_{ij}^{(l)} \tag{3.34}$$

$$\frac{\partial}{\partial b_i^{(l)}} J(W,b) = \frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial b_i^{(l)}} J\left(W,b;x^{(i)},y^{(i)}\right) \tag{3.35}$$

3.4.3 Convolutional Layer

The neural network mentioned earlier is a fully connected network, which means all hidden units are connected to all input units. For image processing, this approach allows feature learning on smaller images. However, for bigger images, if you want to use a fully connected network for feature learning, the amount of computation and parameters required is very large. A simple solution to this problem is to limit the connection between the hidden unit and the input unit. For image processing, each hidden unit will only be connected to a small portion of the contiguous pixel area of the input data.

Natural images have the characteristic of stationarity, that is, one part of the image has the same statistical characteristics as other parts. This shows that the features we learn in one part of the image can also be applied to other parts of this image. More precisely, it is to learn features from small blocks of cells randomly sampled form larger images. The learned features can be convolved in a larger image, getting different feature activation values at each part in the image. This is the role of the convolutional kernel in the convolutional layer[65].

The elements that make up the convolutional kernel all contain a weight value and a bias value. Each neuron in the convolutional layer is connected to multiple neurons in the corresponding region in the previous layer. The size of this region is controlled by the size of the convolutional kernel, and the size of this region is also known as the receptive field. During convolutional calculations, the convolutional kernel regularly sweeps through the input data, multiplying the matrix elements on the input data in the receptive field and superimposing the deviation values. The following Figure 3-9 is a schematic diagram of the convolution operation example.

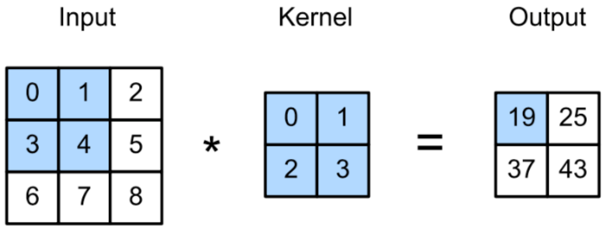


Figure 3-9 The schematic diagram of the convolution operation

The calculation formula for the nth convolutional layer is:

$$h^n = f\left(\sum_{i=1}^{M_x}\sum_{j=1}^{M_y} w_{i,j} \oplus h^{n-1}(i,j) + b^n\right) \tag{3.36}$$

In this formula, M_x and M_y are the length and width of the convolutional kernel M. $w_{i,j}$ is the weight in the convolutional kernel, h^n represents the features of this layer, and h^{n-1} represents the input feature of the previous layer. b^n represents the bias term in the nth layer convolution kernel. $\oplus$ represents the two-dimensional convolution.

3.4.4 Using Deep-CNN for Facial Images Emotion Recognition

Although CNN already has good image classification capabilities, such models are still lack of learning for image spatial invariance. Therefore, a deep-CNN is proposed in this chapter for image classification. This model is a further improvement based on CNN. It improves the pretraining method and proposes new network training steps. Deep-CNN adds the pooling layer in the neural network. This pooling layer enables the learning of local information from the image and allows the model to better adapt to shifts in image parts.

In order to achieve higher accuracy result of emotion recognition, deep-CNN is used in this system. The proposed deep-CNN has four convolutional layers and two fully connected layers. The size of the input image is 48*48. The training set, test set ratio is 7:3, which means it is trained on 18375 samples and tested on 7875 samples. The emotion recognition result includes seven types: happy, sad, surprise, fear, anger, disgust, neutral.

The basic structure of deep-CNN is shown in Figure 3-10 and 3-11. The final accuracy on test set is 88.1%, and the specific result of each emotion recognition can be obtained in Table 3-1. It shows that deep-CNN structure has a good performance on facial images emotion recognition.

Table 3-1 The emotion recognition result of facial images using deep-CNN on seven emotion types.

Table 3-1 The emotion recognition result of facial images using deep-CNN on seven emotion types

Emotion Types	Total Number of Facial Images	Numbers of Testing Images	Accuracy
Happy	3060	918	83%
Sad	600	180	86.7%
Surprise	1470	441	89.8%
Fear	1200	360	86.7%
Anger	1290	387	91.5%
Disgust	750	225	85.3%
Neutral	17880	5364	88.8%
Total	26250	7875	88.1%

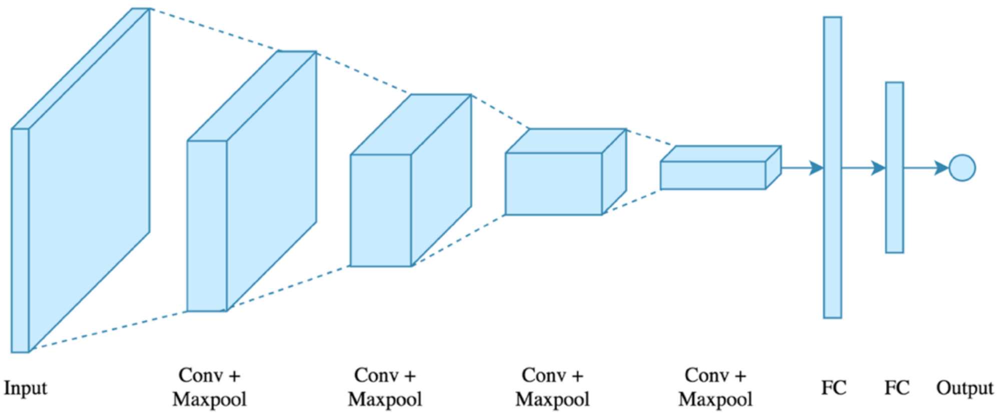


Figure 3-10 The basic structure of the proposed deep-CNN[66]

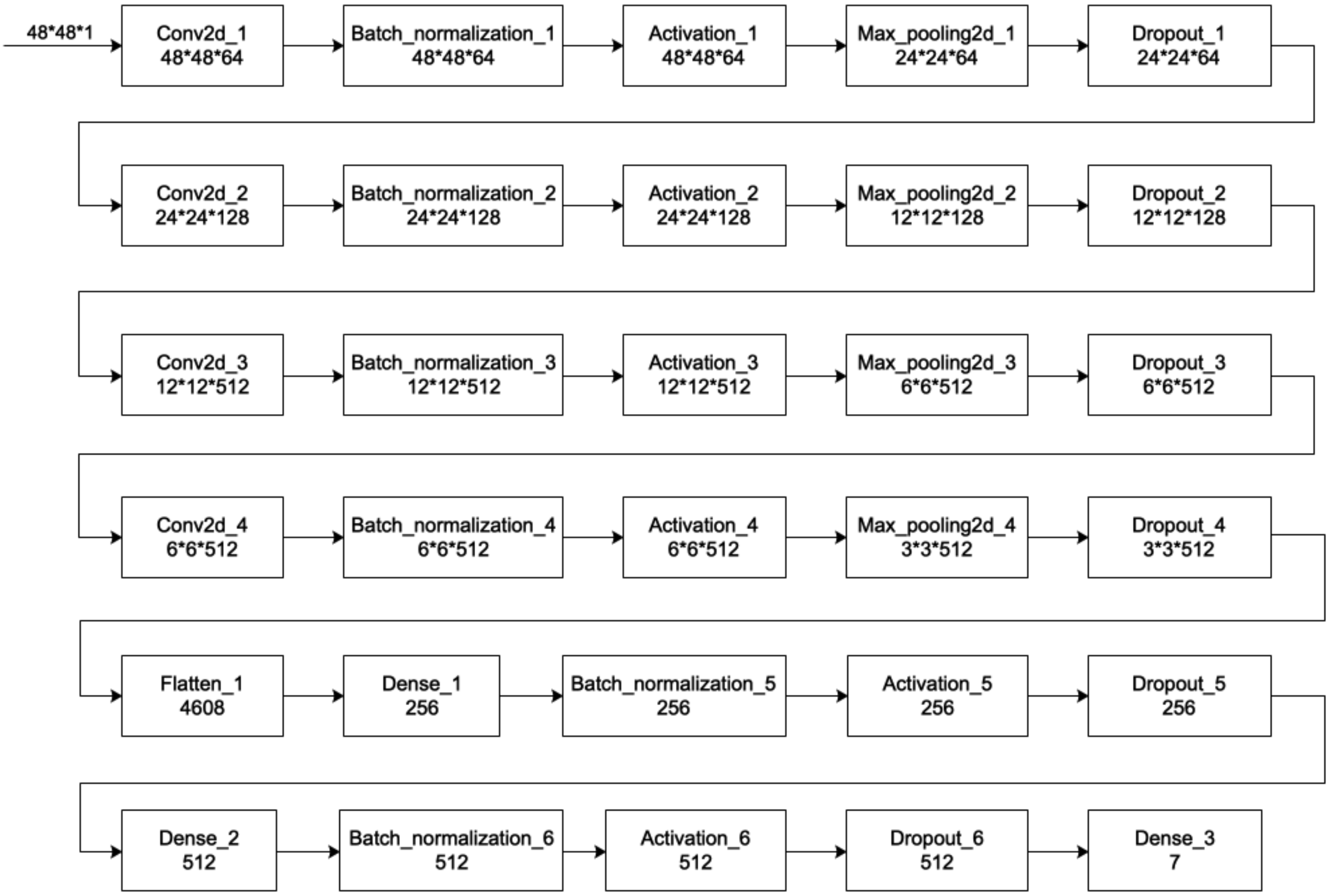


Figure 3-11 The specific internal structure of deep-CNN

3.5 Chapter Summary

The processing procedures for video data in the DEAP dataset are described in this chapter. The facial images are first extracted from the videos. The Adaboost algorithm is used for face detection, and the detected face range of the subject is cropped according to the face detection result. The image of the captured key face is then preprocessed, and the main steps include grayscale processing of the color image and grayscale equalization. The subject's facial preprocessed images are fed into the deep-CNN model for training. The trained deep-CNN model achieved 88.1% accuracy in recognizing emotions.

4 A Dual-Modality System for Data Fusion Based on EEG and Facial Images

4.1 Basic Steps for A Dual-Modality System Data Fusion

This system uses two kinds of physiological data based on the DEAP dataset, including facial video data and EEG data, to do the data fusion on decision-level, which can improve the accuracy of emotion recognition. The Figure 4-1 shows the flow chart of the system for dual-modality emotion recognition.

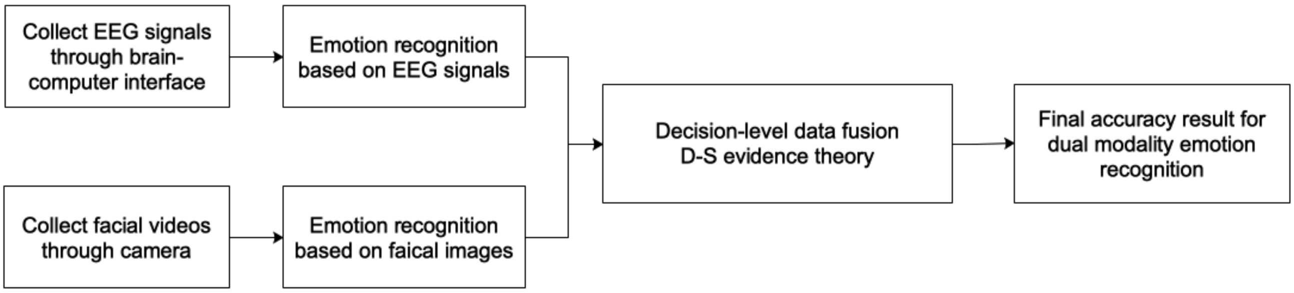

Collect EEG signals through brain-computer interface
Emotion recognition based on EEG signals
Collect facial videos through camera
Emotion recognition based on faical images
Decision-level data fusion D-S evidence theory
Final accuracy result for dual modality emotion recognition

Figure 4-1 The flow chart of dual-modality system

The dataset mainly includes two parts, one part is EEG signals collected through brain-computer interface, and the other part is the facial videos collected through camera. Firstly, preprocess the collected facial video and EEG signals and extract feature values. The emotion quantification calculation is then performed in two modules. Finally, fuse the results from two single modality emotion recognition systems and get the final emotional recognition results.

4.2 Multi-Source Information Fusion Theory

Information fusion[67], also known as data fusion, uses certain fusion rules to synthesize and process information from multiple different channels and sources to obtain final results. Multi-source information fusion algorithm can effectively improve the deficiencies of a single channel, improve the anti-interference ability of the system, increase system stability and wrapping in time and space. The fusion methods mainly are data level fusion, feature level fusion and decision level fusion according to the stage of fusion. The flowchart of each method is shown in Figure 4-2, Figure 4-3 and Figure 4-4.

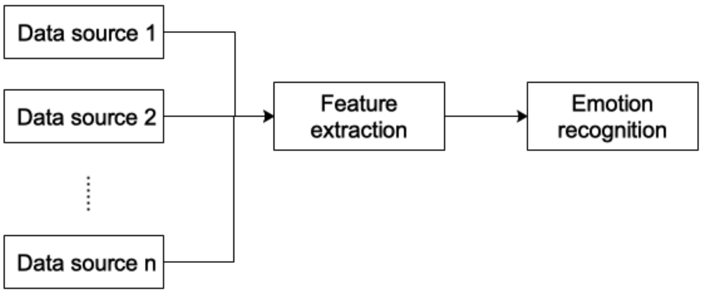


Figure 4-2 The schematic diagram of data-level fusion

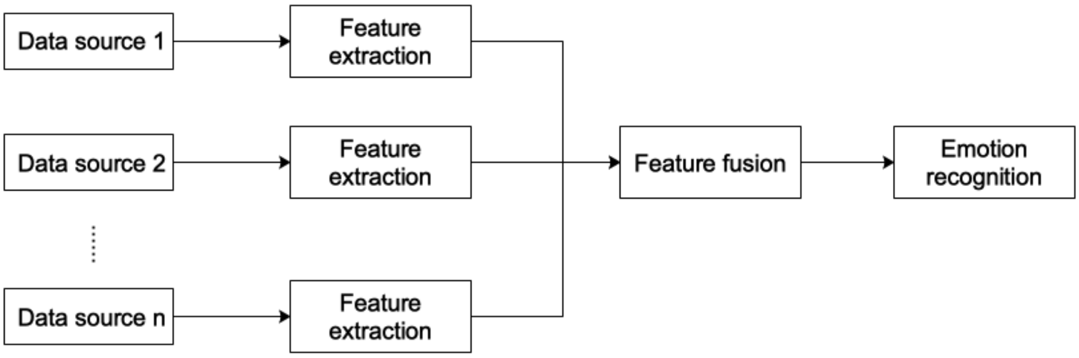


Figure 4-3 The schematic diagram of feature-level fusion

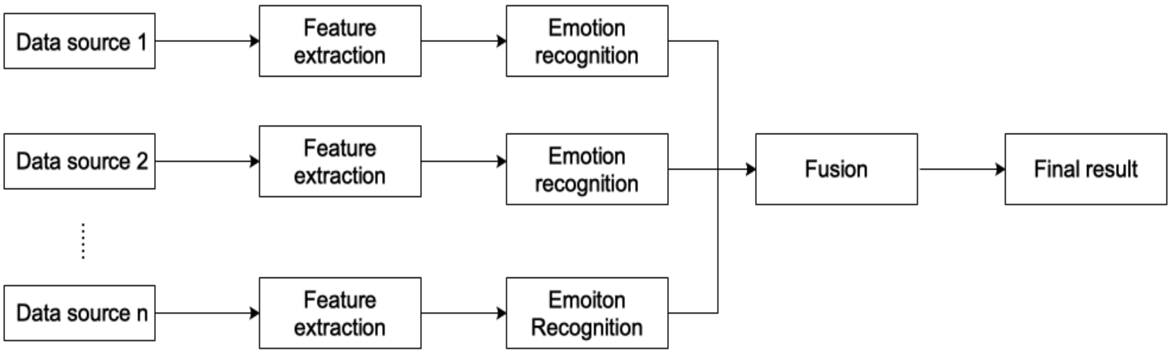


Figure 4-4 The schematic diagram of data-level fusion

The goal of this thesis is to achieve emotion recognition through two kinds of signals: EEG signals and facial images. These two signal sources, EEG source and facial image source, are used as input data for analysis. They belong to different categories and are not suitable for data level fusion. The EEG signals and facial images in this thesis are classified by neural network, so it is not appropriate to do the fusion on feature level. Therefore, decision level fusion is finally used to fuse EEG signals and facial images. The corresponding classifiers are established based on EEG and facial images, and the emotion recognition results from two classifiers are used for decision level fusion.

4.3 D-S Evidence Theory for Decision Level Fusion

Usually, the decision level information fusion methods includes Bayesian[68], neural networks, fuzzy probability theory and D-S evidence theory[69]. The D-S evidence theory is selected in this thesis to fuse facial image and EEG signal emotion classifiers. The recognition framework is:

$$\begin{gathered} F = Happy(F_1), Sad(F_2), Surprise(F_3), \\ Fear(F_4), Anger(F_5), Anger(F_6), Neutral(F_7) \end{gathered} \tag{4.1}$$

$m_i(F_i)$ represents the basic probability of each emotional state, and $F_i \in F$. After calculating the basic probability of facial image emotion and EEG emotion, the final recognition result is obtained using D-S fusion algorithm.

4.3.1 BPA Function

In an n classification problem, the recognition accuracy rate is $\varepsilon_r^{(i)}$, the recognition error rate is $\varepsilon_w^{(i)}$, and the recognition rejection rate is $\varepsilon_\eta^{(i)}$. It is not hard to get:

$$\varepsilon_r^{(i)} + \varepsilon_w^{(i)} + \varepsilon_\eta^{(i)} = 1 \tag{4.2}$$

Assuming that a test sample is classified to be class A_k, the greater number of all training samples classified to class A_k, the more like it is to belong to that class in the real situation. On the contrary, the training sample is more less likely to belong to that class. In order to avoid the impact of classification uncertainty, the error rate of some classifications is generally redistributed and rearranged according to the sample distribution. Its BPA function is as follows:

$$M_i(\Theta) = \varepsilon_{ri}^{(i)} \tag{4.3}$$

$$M_i(A_k) = \varepsilon_r^{(i)} \tag{4.4}$$

$$M_i(A_l) = \frac{vote_i(j,t) \cdot \varepsilon_w^{(i)}}{\sum_{t \subset U, t \neq j} vote_i(j,t)} j \neq k \tag{4.5}$$

where U is the set of all classifications, $vote_i(j,t)$ is the number of samples that are classified in class j in all samples when the sample is classified as class j by the first classifier.

4.3.2 D-S Evidence Theory

Suppose U is the identification frame, $m(F_i)$ is the basic probability distribution function obtained by different evidence, and ϕ is the uncertain set. ε_1 and ε_2 are the two thresholds, which are usually summed up by a large number of experiments. There are four ways for type determination:

1) The basic probability assignment value of the target class is the largest:

$$m(F_1) = max\{m(F_i), F_i \subset U\} \tag{4.6}$$

$$m(F_2) = max\{m(F_i), F_i \subset U \,\&\, F_i \neq F_1\} \tag{4.7}$$

2) The basic probability of the target class is greater than the uncertainty interval:

$$m(F_1) > m(\phi) \tag{4.8}$$

3) The uncertainty interval must be less than a certain threshold:

$$m(\phi) < \varepsilon_2 \tag{4.9}$$

4) The probability assignment value of the target minus the probability assignment value of other classes must be greater than a certain threshold:

$$m(F_1) - m(F_2) > \varepsilon_1 \tag{4.10}$$

Based on the above rules, it can be inferred that F_1 is the classification result.

4.4 A Dual-Modality System for Data Fusion

Facial expression based emotion recognition has now developed quite maturely. Facial images of subjects are collected using cameras and a face detection algorithm is used for cropping facial area. After image preprocessing, more tractable images are obtained and then sent to the trained neural network for emotion recognition. The emotion recognition of a single modality often results in a low recognition accuracy rate due to the characteristics of the modality itself. The multi-modality approach can effectively combine the advantages of different modalities and make up for the shortcomings of a single modality. Therefore, we propose two separate neural networks for EEG emotion recognition and facial image emotion recognition and then use a decision level fusion method to fuse the results. It implements dual-modality emotion recognition and improves accuracy.

There are two ways of classifying emotions. The discrete emotion model divides human emotions into anger, happiness, sadness, surprise, disgust, and fear. The two-dimensional continuous model uses two dimensions of arousal and valence. The classification method based on EEG signals is using two-dimensional continuous emotion model. The emotion classification based on facial images uses the discrete emotion model. Those two models have to be unified to the same model before decision level fusion. DEAP dataset collected EEG data and other physiological data of the subjects when they are watching different video clips, and the first 22 subjects' facial expressions are also recorded. The EEG data has four dimensions of arousal, valence, liking, and dominance, and those four dimensions are specifically quantified between 0 and 9. According to the two-dimensional emotion model, 4.5 is taken as the median value which corresponds to the discrete model, as shown in Figure 4-5.

Considering converting the results of EEG emotion recognition on two dimensional continuous model to the seven basic emotion types. The corresponding emotion results of EEG emotion recognition using SVM-PCA is show in Table 4-1.

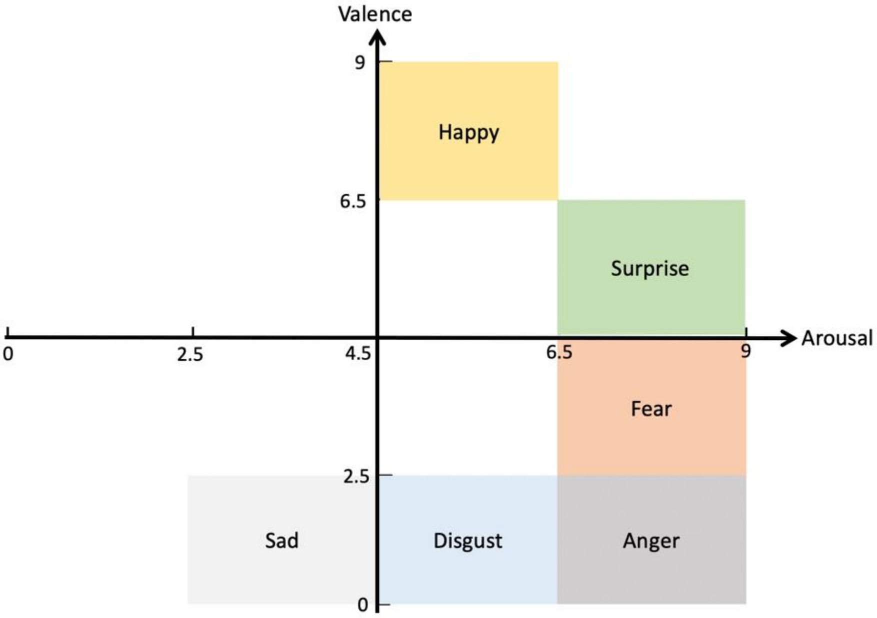


Figure 4-5 The correspondence of seven basic emotions on a two dimensional continuous model

Table 4-1 The emotion recognition accuracy of EEG using SVM-PCA on seven basic emotion types

Emotion Types	Total Number of Data	Number of Testing Data	Accuracy
Happy	3060	918	62.42%
Sad	600	180	65.00%
Surprise	1470	441	60.09%
Fear	1200	360	66.67%
Anger	1290	387	61.50%
Disgust	750	225	64.44%
Neutral	17880	5364	64.95%
Total	26250	7875	64.28%

After the EEG discriminant result and the facial emotion recognition result are unified into the discrete emotion model, the D-S decision fusion is used to combine the two discriminant results and get the final emotion recognition result. Decision-level fusion is directly aimed at specific decision-making goals. Compared with feature-level fusion, the processing cost of decision fusion is relatively low and the anti-interference ability is strong. Common decision fusion methods include Bayesian inference, expert system, D-S evidence theory, fuzzy set theory, etc. Since D-S fusion has the advantages of good convergence and high reliability of additive fusion method, we finally choose D-S evidence decision fusion method, and the final emotion recognition accuracy is 92.3%. The specific accuracy results on seven emotion types are shown in Table 4-2 and Figure 4-6.

Table 4-2 The results of SVM-PCA based on EEG, deep-CNN based on facial images, and dual modality system on emotion recognition

Emotion Types	Accuracy of EEG Using SVM-PCA	Accuracy of Facial Images Using Deep-CNN	Accuracy of Dual-Modality System
Happy	62.42%	83%	93.36%
Sad	65.00%	86.7%	82.78%
Surprise	60.09%	89.8%	93.20%
Fear	66.67%	86.7%	91.67%
Anger	61.50%	91.5%	92.25%
Disgust	64.44%	85.3%	85.78%
Neutral	64.95%	88.8%	92.69%
Total	64.28%	88.1%	92.30%

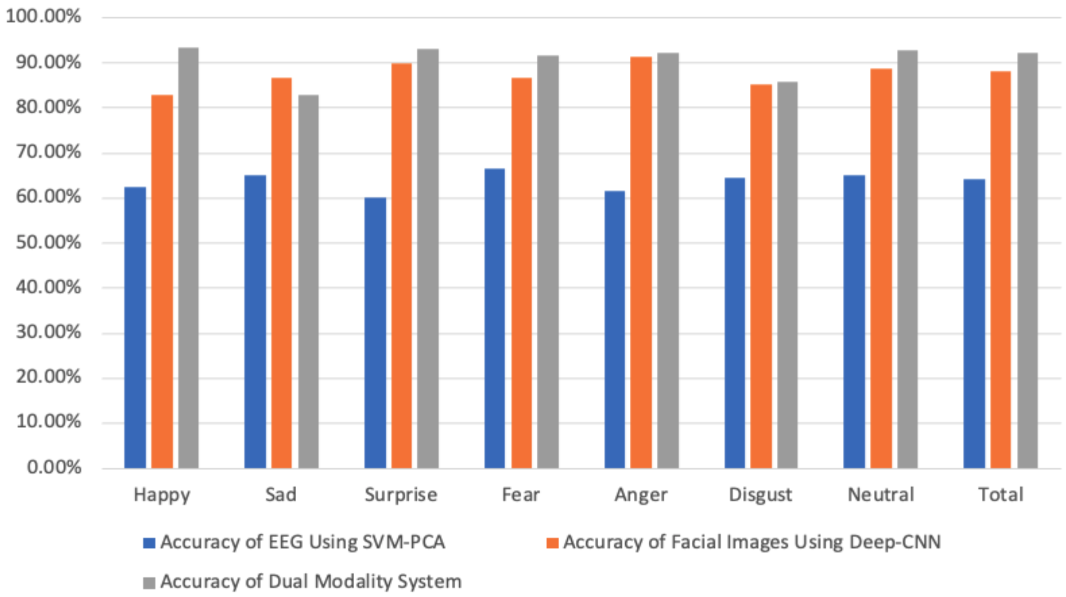


Figure 4-6 The comparison of emotion recognition accuracy between single modality systems and the dual-modality system

It can be seen from the table that the emotion recognition network proposed in this thesis, which fuses the EEG signal emotion recognition and facial expression recognition, has a better accuracy result than the single modality emotion recognition method. It fully demonstrates the superiority of the dual-modality system.

The researches of dual-modality emotion recognition using EEG and facial images on DEAP dataset are not common. Some of the relevant study results are listed in Table 4-3 and Table 4-

4 for comparison. Tabel 4-3 focuses on the researches of EEG emotion recognition, and the dataset used is quite similar with DEAP. Table 4-4 focuses on the dual-modality emotion recogniton systems based on EEG or facial images combing with other physiological signals or speech signals.

Table 4-3 The results of comparison with single modality EEG emotion recognition system

Data Type	Author	Dataset	Accuracy
EEG	KHOSROWABADI R[70]	26 subjects under video stimulus	84.5%
EEG	LEEY[71]	40 subjects under video stimulus	79%
EEG	QING C[72]	DEAP, SEED	DEAP 57.15% SEED 71.41%
EEG	CHEN T[73]	19 subjects under video stimulus	83.34%
EEG + Facial Images	Chapter 4	DEAP	92.3%

Table 4-4 The results of comparison with dual-modality emotion recognition system

Data Type	Author	Dataset	Accuracy
EEG + PPS	Koelstra[74]	DEAP	61.5%
EEG + PPS	Tang[75]	DEAP	83.5%
EEG + PPS	Yin[76]	DEAP	83.5%
Speech + Facial Images	Nguyen[77]	eNTER FACE’05	90.85%
Speech +Facial Images	Zhang[78]	eNTER FACE’05	85.97%
EEG + Facial Images	Chapter 4	DEAP	92.3%

4.4 Chapter Summary

This chapter begins with an introduction to the basic steps of a dual-modality system. Then, the principle and types of multi-source information fusion theory are introduced, including data-level fusion, feature-level fusion, and decision-level fusion. Subsequently, the relevant theoretical basis of D-S fusion at decision-level is elaborated. The dual-modality system can compensate for the shortcomings of single modality systems, such as a lack of data information.

The system's robustness and trustworthiness can be increased after dual-modality fusion. The results of the network structure in Chapters 2 and 3 are fused at the decision-level using the DEAP dataset, according to the D-S evidence theory. The emotion recognition accuracy is improved compared to the single modality system.

5 Application of the Proposed Dual-Modality Emotion Recognition System in Educational Scene

5.1 The Significance of Emotion Recognition in Educational Scenes

With the rapid development and widespread popularization of information technology, e-learning has become a normal form of all kinds of education. Especially in recent years, due to the impact of the covid-19, more and more schools and educational institutions have adopted the form of online teaching.

Compared with offline teaching, online teaching has higher requirements for teachers' teaching content and methods. It is difficult for teachers of online teaching to adjust the pace and methods directly by receiving emotional feedback from students, so they need to explore and design ways that can improve the effectiveness of teaching[79]. The dual-modality emotion recognition system proposed in this thesis can be used to guide the teaching design and better dataize the students' emotional state.

5.2 Experiment Design and Data Collection

Experiments in the educational scene can be designed to test the effects of the dual-modality emotion recognition system proposed in this thesis. The process of this experiment is as follows:

1) Recruit 20 healthy participants (half male and half female) aged between 20 and 30 years old.

2) Prepare 40 different content teaching videos, and each video lasts for one minute.

3) Prepare the EEG acquisition equipment, cameras, and other equipment for data acquisition during the experiment.

4) Prepare a quiet laboratory for an emotion recognition experiment. Before the start of the experiment, the subject needs to be informed of the experimental process and precautions, and the experiment begins when the subject is ready.

5) Place the electrodes on the subject’s head according to the international 10-20 standard. The camera is facing the subject, ensuring that the subject’s facial expression can be recorded completely. The computer screen used for playing teaching videos is positioned in front of the subject. In order to ensure good light conditions, all experiments are carried out under the condition of turning on the laboratory lights.

6) The subject is first asked to sit quietly for 5 minutes. Collect the EEG signals and the facial expression videos within 5 minutes as a reference. Stimulating videos are then played on the

computer screen and data is collected during watching.

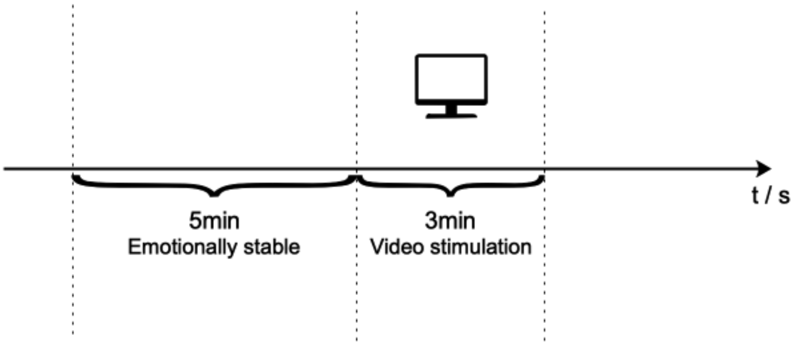
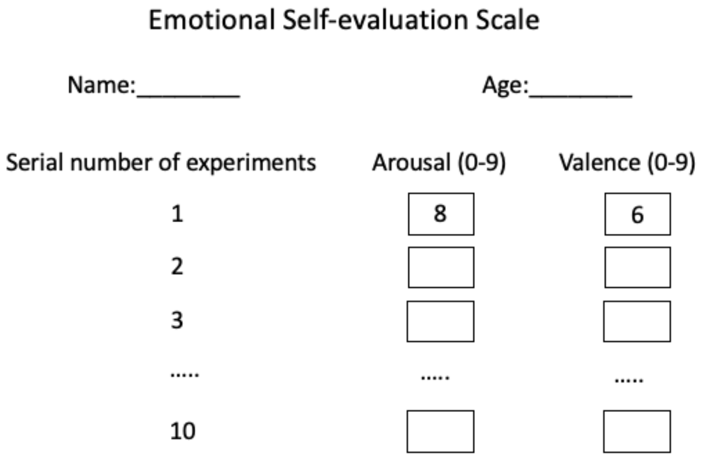


Figure 5-1 Schematic diagram of the data collection experiment

At the end of each video stimulation, subjects are required to fill out an emotion self-evaluation scale. The scale includes personal information such as name, age, and the values of valence and arousal for each experiment, ranging from 0 to 9. The emotional self-evaluation scale is shown in Figure 5-2.



Emotional Self-evaluation Scale

Name:________ Age:________

Serial number of experiments	Arousal (0-9)	Valence (0-9)
1	8	6
2		
3		
.....	.....	.....
10		

Figure 5-2 An example of emotional self-evaluation scale

For the EEG data and facial video data of 20 subjects collected, those data is divided into training set and test set. The training dataset is for training the dual-modality system and the test data is for testing the emotion recognition accuracy.

5.3 Experiment Verification on the Proposed Dual-Modality System

5.3.1 Data Acquisition

Based on the experimental design method proposed above, one subject is selected as the experimental data acquisition object. Before the experiment, 40 one-minute online teaching videos are prepared, which is the duration of a typical whole class. The main equipment used in the experiment is a convenient EEG acquisition device, laptop, Android mobile phone. The picture of the EEG acquisition device and the diagram of wearing the device is shown in Figure 5-3. The laptop is used for playing teaching video clips and recording the subject’s facial expressions. The Android phone is for connecting with the EEG acquisition device through Bluetooth. The interface between the mobile phone and EEG acquisition device is shown in Figure 5-4.

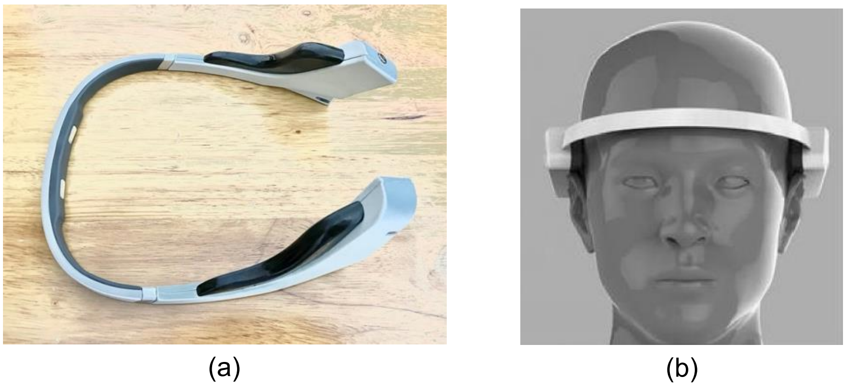
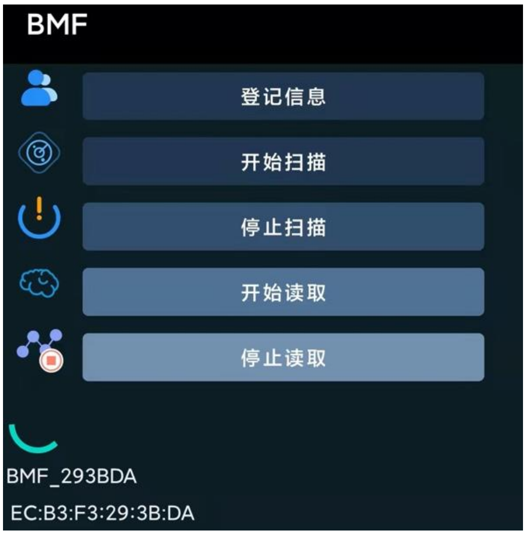


Figure 5-3 (a) The picture of EEG acquisition device (b) The diagram of wearing EEG acquisition device



BMF_293BDA
EC:B3:F3:29:3B:DA

Figure 5-4 The interface between the mobile phone and EEG acquisition device (From top to bottom, the buttons are for subject’s information registration, start scaning, stop scanning, start loading data and stop loading data)

The subject is asked to wear the EEG acquisition device for testing before the experiment starts. After the subject is familiar with the experiment steps, then begin the actual experiment. 40 videos of the subject's facial expressions are recorded with .mov format and each of the videos lasts for one minute. What's more, the EEG signals are uploaded to the server during the experiment and are used for further analysis. After each small experiment, the subject is required to fill the form in Figure 5-2.

5.3.2 Evaluation of Teaching Effectiveness

The collected EEG waveforms are shown in the app of the mobile phone in real-time. Figure 5-5 shows the different EEG frequency bands waveform of the subjects during the experiment. For the three columns in Figure 5-5, the first one "NO noise" means the EEG acquisition device is worn correctly. The second one "Attention" and the third one "Meditation" are numbers calculated from EEG signals using the FFT method, which shows the subject's current state of concentration in learning.

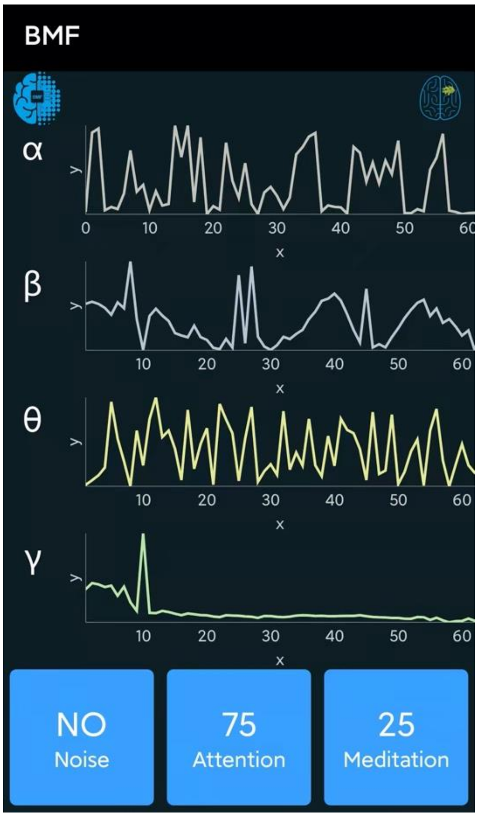


Figure 5-5 Real-time EEG acquisition interface on Android phone

Considering seven types of emotion classification are not easy to highlight the learning status of students in actual educational scene, we change the emotion types into four typical emotions: happy, relaxed, sad and fear. The correspondence between those four emotion types and two dimensional continuous emotion is shown in Figure 5-6. In Figure 5-7, Figure 5-8 and Figure 5-9, representative examples of four emotion types during the experiment are listed.

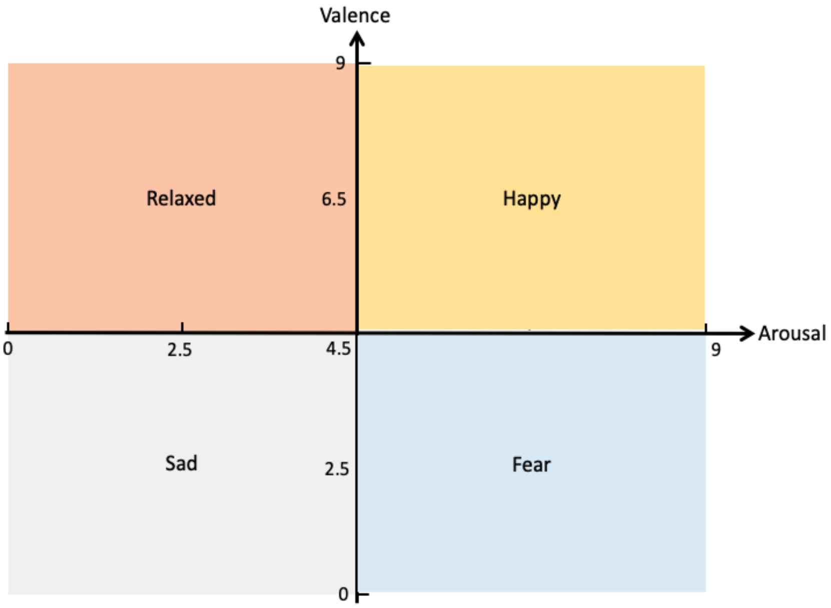


Figure 5-6 The four discrete emotion classifications used in educational scene

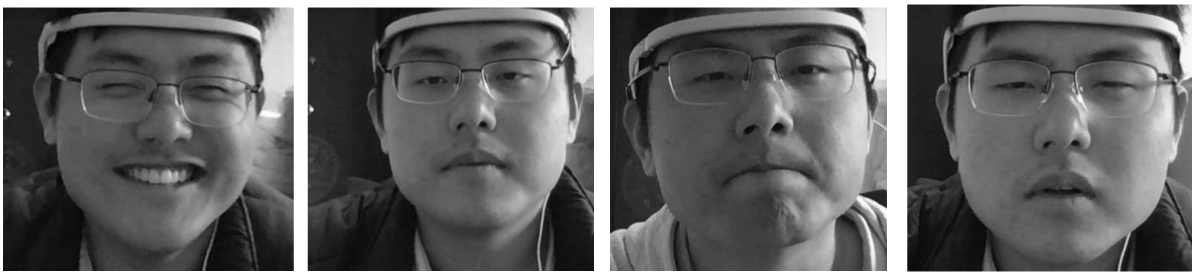


Figure 5-7 The four basic emotion states of the subject during experiment (from left to right, respectively, happy, relaxed, sad, fear)

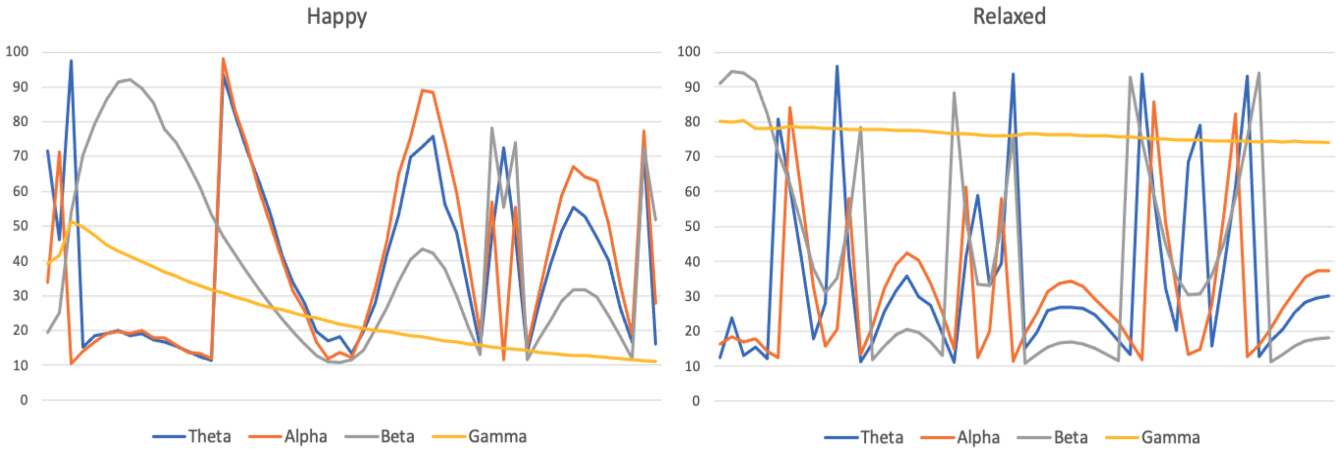


Figure 5-8 The EEG signal of the subject in emotion “happy” and “relaxed” during experiment

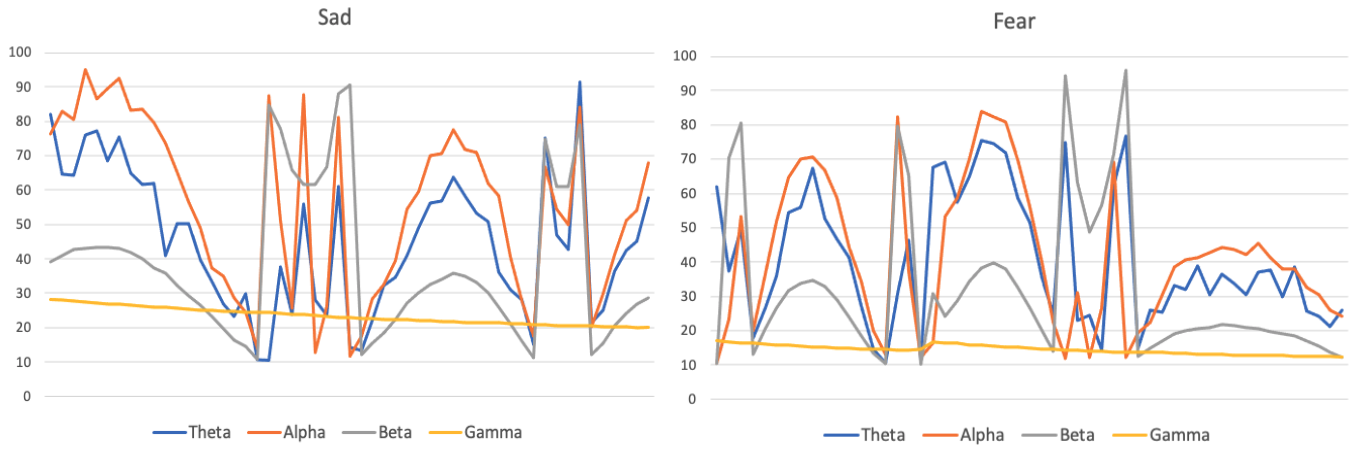


Figure 5-9 The EEG signal of the subjects in emotion “sad” and “fear” during experiment

In Figure 5-8 and Figure 5-9, the blue line represents the Theta band, the red line represents the Alpha band, the gray line represents the Beta band, and the yellow line represents the Gamma band. The horizontal axis means EEG sampling points within a one-minute video. The vertical axis means the amplitude of each frequency band.

The collected video of the subject’s facial expression is totally 40 clips. First, one image is captured every 50 video frames and a total of 1400 photos are obtained. The 1400 photos are then preprocessed, including face detection, image cropping, grayscale equalization, and feature extraction. Those processed images are then fed into the deep-CNN neural network for training and testing. EEG signals are analyzed with the SVM-PCA method. The results of two single modality systems are fused at the decision-level using D-S evident theory system. The final emotion recognition accuracy of the collected data using dual-modality system is 82.3%.

Table 5-1 The emotion recognition result of the subject using the proposed dual-modality system

Emotion Types	Total Number of Data	Numbers of Testing Data	Accuracy of Dual Modality System
Happy	385	115	86.09%
Sad	175	52	82.69%
Fear	315	94	82.98%
Relaxed	525	157	78.98%
Total	1400	418	82.30%

Teachers can use the result of dual-modality emotion recognition system to understand the emotional state of students during class. If the students are generally in a low emotional state during a certain period, the teaching methods for this period can be appropriately adjusted. If the student's emotions are relatively high during a certain period or the narration of a certain knowledge point, it means that the teaching method has universal applicability to students.

The emotion recognition accuracy using data collected real online teaching experiment is not as good as the DEAP dataset, and this may be caused by several reasons:

1) The amount of dataset is small compared with the DEAP dataset. In DEAP, there are 22 subjects and each of the subjects watches 40 stimulus video clips. In the online teaching experiment, there is only one subject. An insufficient number of subjects resulted in a small amount of collected data.

2) The types of video stimuli are different. In DEAP, videos with more pronounced emotional tendencies are specifically selected. However, in a real teaching scene, the whole teaching process may not be so uplifting or irritative and is more likely to be gradual and lively.

3) The status of subjects is different. In DEAP, subjects are required to focus on the videos. On the contrary, in the educational scene, some mind wandering phenomena may exist. The emotion recognition result under this kind of status may be incorrect.

5.4 Chapter Summary

This chapter describes the possibilities and applications of dual-modality emotion recognition systems for the educational scene. A specific experimental process is designed, in which the actual EEG and face video data of the subjects are collected during the experiment and used for training the model. Finally, the results of the emotion classification of the subjects and the accuracy of the model are obtained. This model can be used to explore and improve the methods teachers use in class.

6 Summary and Outlook

6.1 Summary

With the continuous development of online learning, online live courses and pre-recorded courses con replace offline teachers to complete more and more work. Achieving better interaction between teachers and students is an important aspect. In order to get better teaching effects, teachers should be able to actively adjust teaching methods from students' emotional feedback in class. Therefore, the main purpose of this thesis is to design a system that can perform emotion recognition and can be applied in the educational scene.

There are two main methods of identifying emotions, one is based on signals such as face and speech, this method is simple and fast[80]. But there will be a problem that the recognition rate is not high caused by the subject deliberately concealing emotions. The other is the method based on physiological signals such as EEG and EMG[81], which has a more complex collecting process. It can avoid misjudgment caused by concealing emotions under subjective factors.

This thesis mainly studies the deep learning system on EEG signals and facial expressions. The network structure is improved in the single modality emotion recognition systems of EEG and facial expression. The recognition rate has been improved, while the number of parameters is reduced, making the training process easier and faster. The two neural networks are combined into one dual-modality system in decision-level fusion for emotion recognition. The main work completed in this thesis are as follows:

The main research content of this thesis is determined through reading relevant literature at home and abroad. The research background and significance are also introduced. Subsequently, the current research status of emotion recognition based on EEG signals and facial images is described. Some basic emotion recognition steps and methods are also included.

An emotion recognition method of EEG signals based on SVM-PCA is proposed. It introduces the generation principle, frequency band characteristics, emotion classification theory, and DEAP dataset based on EEG signals. It also includes an introduction to the DEAP dataset used in the thesis. This part focuses on the filtering and preprocess methods of EEG signals, such as ocular artifact removal, decomposition, and feature extraction using wavelet transforms. Finally, the SVM-PCA network is trained using EEG signals collected from subjects, and the accuracy of this network is obtained for subsequent analysis.

A facial expression emotion recognition algorithm based on deep-CNN is proposed. In this part, firstly get the screenshots of the videos. Those screenshots are processes for face detection and image capture using the Adaboost algorithm. Grayscale equalization and feature extraction are then performed on the captured images. Then the images are fed into the deep-CNN for training, and we can get the emotion classification results and the accuracy of this neural network.

It gives the basic fusion framework of EEG emotion recognition using SVM-PCA and facial expression emotion recognition using deep-CNN. The decision-making model of multivariate information fusion theory and D-S evidence theory is introduced. The two kinds of data are used to identify emotions, and they are fused at the decision level. The recognition rate reaches more than 90%. It is proved that the recognition accuracy of the dual-modality system is higher than that of the single modality model, which also means the dual-modality system can overcome the shortcomings of the single modality system.

Based on the dual-modality emotion recognition system mentioned in this article, the data collection method is designed for teaching scenarios. The specific process and data collection method of the experiment are introduced. The EEG and facial image data of the subject is collected using the designed method. The data is then analyzed using the proposed dual-modality system for emotion recognition. The results obtained can be used to evaluate the actual teaching effect.

6.2 Outlooks

This thesis mainly aims at the content of deep learning in emotion recognition and proposed a dual-modality emotion recognition system that can be used in educational scenes. The system is based on EEG signals and facial expression images collected from subjects, and the two kinds of data are fed into a neural network for training and classification. Finally, the results from the two classifiers are fused at the decision level. Compared with single modality systems, the recognition rate of the dual-modality system has been improved, but there are still some shortcomings in the system:

From the experimental results, it can be found that the emotion recognition accuracy of EEG signals based on SVM-PCA is not high enough. In future research, we could improve the accuracy from denoising, feature selection, network structure optimization, etc.

For decision-level fusion, there are many ways to do it. Only one of these methods has been selected in this thesis. In subsequent studies, more methods can be considered to do the data fusion and find the most effective one.

Due to the limitations of time, experimental equipment, experimental resources, etc., it is quite difficult to recruit enough subjects for experiments. For further application in educational scene, it could be considered increasing the number of experiments to expand the dataset.

Reference

[1] Ferretti V, Papaleo F. Understanding others: Emotion recognition in humans and other animals[J]. Genes, Brain and Behavior, 2019, 18(1): e12544.

[2] Zeng Z, Pantic M, Roisman G I, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 31(1): 39-58.

[3] Pan X. Research on the emotion recognition based on the fuzzy neural network in the intelligence education system[C]//2011 Second International Conference on Digital Manufacturing & Automation. IEEE, 2011: 1030-1033.

[4] Kessous L, Castellano G, Caridakis G. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis[J]. Journal on Multimodal User Interfaces, 2010, 3(1): 33-48.

[5] Minaee S, Minaei M, Abdolrashidi A. Deep-emotion: Facial expression recognition using attentional convolutional network[J]. Sensors, 2021, 21(9): 3046.

[6] Shaheen S, El-Hajj W, Hajj H, et al. Emotion recognition from text based on automatically generated rules[C]//2014 IEEE International Conference on Data Mining Workshop. IEEE, 2014: 383-392.

[7] Karna M, Juliet D S, Joy R C. Deep learning based text emotion recognition for chatbot applications[C]//2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184). IEEE, 2020: 988-993.

[8] Shen Z, Cheng J, Hu X, et al. Emotion recognition based on multi-view body gestures[C]//2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 3317-3321.

[9] Wei G, Jian L, Mo S. Multimodal (Audio, Facial and Gesture) based Emotion Recognition challenge[C]//2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE, 2020: 908-911.

[10] Anusha R, Subhashini P, Jyothi D, et al. Speech emotion recognition using machine learning[C]//2021 5th International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2021: 1608-1612.

[11] Kim J, André E. Emotion recognition based on physiological changes in music listening[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 30(12): 2067-2083.

[12] Schleicher R, Antons J N. Evoking emotions and evaluating emotional impact[M]//Quality of Experience. Springer, Cham, 2014: 121-132.

[13] Picard R W, Vyzas E, Healey J. Toward machine emotional intelligence: Analysis of affective physiological state[J]. IEEE transactions on pattern analysis and machine intelligence, 2001, 23(10): 1175-1191.

[14] Al-Nafjan A, Hosny M, Al-Ohali Y, et al. Review and classification of emotion recognition based on EEG brain-computer interface system research: a systematic review[J]. Applied Sciences, 2017, 7(12): 1239.

[15] Shu L, Xie J, Yang M, et al. A review of emotion recognition using physiological signals[J]. Sensors, 2018, 18(7): 2074.

[16] Olejniczak P. Neurophysiologic basis of EEG[J]. Journal of clinical neurophysiology, 2006, 23(3): 186-189.

[17] Al-Nafjan A, Hosny M, Al-Ohali Y, et al. Review and classification of emotion recognition based on EEG brain-computer interface system research: a systematic review[J]. Applied Sciences, 2017, 7(12): 1239.

[18] Homan R W, Herman J, Purdy P. Cerebral location of international 10–20 system electrode placement[J]. Electroencephalography and clinical neurophysiology, 1987, 66(4): 376-382.

[19] Bartels G, Shi L C, Lu B L. Automatic artifact removal from EEG-a mixed approach based on double blind source separation and support vector machine[C]//2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. IEEE, 2010: 5383-5386.

[20] Hosni S M, Gadallah M E, Bahgat S F, et al. Classification of EEG signals using different feature extraction techniques for mental-task BCI[C]//2007 International Conference on Computer Engineering & Systems. IEEE, 2007: 220-226.

[21] Roy Y, Banville H, Albuquerque I, et al. Deep learning-based electroencephalography analysis: a systematic review[J]. Journal of neural engineering, 2019, 16(5): 051001.

[22] Ekman P, Friesen W V. Constants across cultures in the face and emotion[J]. Journal of personality and social psychology, 1971, 17(2): 124.

[23] Mellouk W, Handouzi W. Facial emotion recognition using deep learning: review and insights[J]. Procedia Computer Science, 2020, 175: 689-694.

[24] Simard P Y, Steinkraus D, Platt J C. Best practices for convolutional neural networks applied to visual document analysis[C]//Icdar. 2003, 3(2003).

[25] Lopes A T, De Aguiar E, De Souza A F, et al. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order[J]. Pattern recognition, 2017, 61: 610-628.

[26] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[27] Martinez B, Valstar M F. Advances, challenges, and opportunities in automatic facial expression recognition[M]//Advances in face detection and facial image analysis. Springer, Cham, 2016: 63-100.

[28] Baillet S, Mosher J C, Leahy R M. Electromagnetic brain mapping[J]. IEEE Signal processing magazine, 2001, 18(6): 14-30.

[29] Farrelly, Colleen. (2017). The Neurobiology of Addiction. 10.13140/RG.2.2.18741.37609.

[30] Kostyunina M B, Kulikov M A. Frequency characteristics of EEG spectra in the emotions[J]. Neuroscience and Behavioral Physiology, 1996, 26(4): 340-343.

[31] Newson J J, Thiagarajan T C. EEG frequency bands in psychiatric disorders: a review of resting state studies[J]. Frontiers in human neuroscience, 2019, 12: 521.

[32] Medithe J W C, Nelakuditi U R. Study of normal and abnormal EEG[C]//2016 3rd International conference on advanced computing and communication systems (ICACCS). IEEE, 2016, 1: 1-4.

[33] Van den Broek E L. Ubiquitous emotion-aware computing[J]. Personal and Ubiquitous Computing, 2013, 17(1): 53-67.

[34] Lang P J. The emotion probe: Studies of motivation and attention[J]. American psychologist, 1995, 50(5): 372.

[35] Kuppens P, Tuerlinckx F, Russell J A, et al. The relation between valence and arousal in subjective experience[J]. Psychological bulletin, 2013, 139(4): 917.

[36] S. Koelstra et al., "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," in IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18-31, Jan.-March 2012, doi: 10.1109/T-AFFC.2011.15.

[37] Bird J J, Ekart A, Buckingham C D, et al. Mental emotional sentiment classification with an eeg-based brain-machine interface[C]//Proceedings of theInternational Conference on Digital Image and Signal Processing (DISP’19). 2019.

[38] Zhao G, Liu Y J, Shi Y. Real-time assessment of the cross-task mental workload using physiological measures during anomaly detection[J]. IEEE Transactions on Human-Machine Systems, 2018, 48(2): 149-160.

[39] Croft R J, Barry R J. Removal of ocular artifact from the EEG: a review[J]. Neurophysiologie Clinique/Clinical Neurophysiology, 2000, 30(1): 5-19.

[40] Satpathy R B, Ramesh G P. An EEG atomized artefact removal algorithm: a review[J]. Micro-Electronics and Telecommunication Engineering, 2022: 805-816.

[41] Judith A M, Priya S B, Mahendran R K. Artifact removal from EEG signals using regenerative multi-dimensional singular value decomposition and independent component analysis[J]. Biomedical Signal Processing and Control, 2022, 74: 103452.

[42] Jung T P, Humphries C, Lee T W, et al. Extended ICA removes artifacts from electroencephalographic recordings[J]. Advances in neural information processing systems, 1997, 10.

[43] Jung T P, Makeig S, Humphries C, et al. Removing electroencephalographic artifacts by blind source separation[J]. Psychophysiology, 2000, 37(2): 163-178.

[44] Choi S, Cichocki A, Park H M, et al. Blind source separation and independent component analysis: A review[J]. Neural Information Processing-Letters and Reviews, 2005, 6(1): 1-57.

[45] Pahuja S K, Veer K. Recent approaches on classification and feature extraction of EEG signal: A Review[J]. Robotica, 2022: 1-25.

[46] Candra H, Yuwono M, Handojoseno A, et al. Recognizing emotions from EEG subbands using wavelet analysis[C]//2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, 2015: 6030-6033.

[47] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods[M]. Cambridge university press, 2000.

[48] Gao Q, Yang Y, Kang Q, et al. EEG-based emotion recognition with feature fusion networks[J]. International Journal of Machine Learning and Cybernetics, 2022, 13(2): 421-429.

[49] Cheng C T, Feng Z K, Niu W J, et al. Heuristic methods for reservoir monthly inflow forecasting: A case study of Xinfengjiang Reservoir in Pearl River, China[J]. Water, 2015, 7(8): 4477-4495.

[50] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. Ieee, 2001, 1: I-I.

[51] Mita T, Kaneko T, Hori O. Joint haar-like features for face detection[C]//Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. IEEE, 2005, 2: 1619-1626.

[52] Papageorgiou C P, Oren M, Poggio T. A general framework for object detection[C]//Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271). IEEE, 1998: 555-562.

[53] Karishma A, Krishnan K A, Kiran A, et al. Smart office surveillance robot using face recognition[J]. International Journal of Mechanical and Production Engineering Research and Development, 2018, 8(3): 725-734.

[54] Wu H, Cao Y, Wei H, et al. Face recognition based on Haar like and Euclidean distance[C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 1813(1): 012036.

[55] Ou Z, Tang X, Su T, et al. Cascade AdaBoost classifiers with stage optimization for face detection[C]//International Conference on Biometrics. Springer, Berlin, Heidelberg, 2006: 121-128.

[56] Žeger I, Grgic S, Vuković J, et al. Grayscale image colorization methods: overview and evaluation[J]. IEEE Access, 2021.

[57] Parker J A, Kenyon R V, Troxel D E. Comparison of interpolating methods for image resampling[J]. IEEE Transactions on medical imaging, 1983, 2(1): 31-39.

[58] Kirkland E J. Bilinear interpolation[M]//Advanced Computing in Electron Microscopy. Springer, Boston, MA, 2010: 261-263.

[59] Sonker D, Parsai M P. Comparison of histogram equalization techniques for image enhancement of grayscale images of dawn and dusk[J]. International Journal of Modern Engineering Research (IJMER), 2013, 3(4): 2476-2480.

[60] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity[J]. The bulletin of mathematical biophysics, 1943, 5(4): 115-133.

[61] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain[J]. Psychological review, 1958, 65(6): 386.

[62] Rumelhart D E, Hinton G E, Williams R J. Learning internal representations by error propagation[R]. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[63] Lazovskaya T, Tarkhov D. Multilayer neural network models based on grid methods[C]//IOP conference series: materials science and engineering. IOP Publishing, 2016, 158(1): 012061.

[64] Rumelhart D E, Durbin R, Golden R, et al. Backpropagation: The basic theory[J]. Backpropagation: Theory, architectures and applications, 1995: 1-34.

[65] Albawi S, Mohammed T A, Al-Zawi S. Understanding of a convolutional neural network[C]//2017 international conference on engineering and technology (ICET). Ieee, 2017: 1-6.

[66] Yoon H J, Jeong Y J, Kang D Y, et al. Effect of Data Augmentation of F-18-Florbetaben Positron-Emission Tomography Images by Using Deep Learning Convolutional Neural Network Architecture for Amyloid Positive Patients[J]. Journal of the Korean Physical Society, 2019, 75(8): 597-604.

[67] Xu W, Yu J. A novel approach to information fusion in multi-source datasets: a granular computing viewpoint[J]. Information sciences, 2017, 378: 410-423.

[68] Zhang H K, Huang B. A new look at image fusion methods from a Bayesian perspective[J]. Remote sensing, 2015, 7(6): 6828-6861.

[69] Zhang W, Ji X, Yang Y, et al. Data fusion method based on improved DS evidence theory[C]//2018 IEEE international conference on big data and smart computing (BigComp). IEEE, 2018: 760-766.

[70] Khosrowabadi R, Quek H C, Wahab A, et al. EEG-based emotion recognition using self-organizing map for boundary detection[C]//2010 20th International Conference on Pattern Recognition. IEEE, 2010: 4242-4245.

[71] Lee Y Y, Hsieh S. Classifying different emotional states by means of EEG-based functional connectivity patterns[J]. PloS one, 2014, 9(4): e95415.

[72] Qing C, Qiao R, Xu X, et al. Interpretable emotion recognition using EEG signals[J]. IEEE Access, 2019, 7: 94160-94170.

[73] Chen T, Ju S, Yuan X, et al. Emotion recognition using empirical mode decomposition and approximation entropy[J]. Computers & Electrical Engineering, 2018, 72: 383-392.

[74] KOELSTRA S, MUHL C, SOLEYMANI M, et al. Deap: A database for emotion analysis; using physiological signals[J]. IEEE transactions on affective computing, 2011, 3(1): 18-31.

[75] TANG H, LIU W, ZHENG W, et al. Multimodal emotion recognition using deep neural networks[C]//Poceedings of the International Conference on Neural Information Processing. Guangzhou, China, 2017: 811−819.

[76] YIN Z, ZHAO M, WANG Y, et al. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model[J]. Computer methods and programs in biomedicine, 2017, 140: 93-110. DOI:10.1016/j.cmpb.2016.12.005

[77] Nguyen D, Nguyen K, Sridharan S, et al. Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition[J]. Computer Vision and Image Understanding, 2018, 174: 33-42.

[78] Zhang S, Zhang S, Huang T, et al. Learning affective features with a hybrid deep model for audio–visual emotion recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(10): 3030-3043.

[79] Lasri I, Solh A R, El Belkacemi M. Facial emotion recognition of students using convolutional neural network[C]//2019 third international conference on intelligent computing in data sciences (ICDS). IEEE, 2019: 1-6.

[80] Busso C, Deng Z, Yildirim S, et al. Analysis of emotion recognition using facial expressions, speech and multimodal information[C]//Proceedings of the 6th international conference on Multimodal interfaces. 2004: 205-211.

[81] Yang F, Zhao X, Jiang W, et al. Multi-method Fusion of Cross-Subject Emotion Recognition Based on High-Dimensional EEG[J]. Frontiers in Computational Neuroscience–Editors' Pick 2021, 2022.