A Lightweight Dual-Channel Multimodal Emotion Recognition Network Using Facial Expressions and Eye Movements

Mengcheng Ji[1] (✉), Fulan Fan[2], Xin Nie[3] and Yahong Li[2]

1 College of Computer Science, South-Central Minzu University, Wuhan 430074, China
2 School of Education, South-Central Minzu University, Wuhan 430074, China
3 School of Computer Science &Technology, Huazhong University of Science and Technology, Wuhan 430074, China
2022120373@mail.scuec.edu.cn

Abstract. Emotional understanding plays a crucial role in various fields related to human–computer interaction, emotional computing, and human behavior analysis. However, traditional single-modal methods often struggle to capture the complexity and subtleties of emotional states. With the advances in eye-tracking technology and facial expression recognition technology, eye-tracking and facial expressions provide complementary insight. We combine eye-tracking and facial expressions to conduct emotional research. Combining these two types of information more comprehensively and accurately describes the emotional experience of individuals and improves upon methods using a single mode. Because human emotional changes require event induction, the events and methods of emotion induction are extremely important. We also present a data collection experiment using emotion theory in psychology. We selected three types of emotion-activating images (positive, neutral, and negative) from the Chinese Affective Picture System (CAPS). We design a system to extract features from the collected data, fusing the multi-modal eye tracking and facial expressions. This system is our proposed dual-channel multi-modal emotion recognition lightweight network VGG-inspired LightNet using a convolutional neural network (CNN). This model achieved an accuracy rate of 96.25% in tests using our gathered data. Compared with single-modal emotion recognition methods, combining eye movement signals with facial features improves the accuracy and robustness of recognizing emotional states.

Keywords: Multimodal; Facial expressions; Eye-tracking; Feature fusion; Emotional recognition.

1 Introduction

Emotions play a crucial role in shaping human interactions and experiences. In today's digital age, the proliferation of social media platforms and advances in human-computer interaction technologies have triggered a growing interest in emotion recognition techniques across different disciplines such as computer science, artificial intelligence,

psychology, and education [1]. The vast amount of data generated on online platforms provides a unique opportunity to gain insights into user emotions, which is crucial for enhancing the intelligence of computer systems and improving the quality of human-computer interactions. Despite the growing interest in emotion recognition, existing approaches still face challenges such as low accuracy, inconsistent performance, and susceptibility to environmental noise. The popularity of deep learning has made multimodal emotion recognition a research focus in emotion classification both domestically and internationally [2]. The rise of deep learning techniques has driven research into multimodal emotion recognition, which integrates different behavioral expressions to provide a more comprehensive understanding of emotional states. By exploiting physiological neural states and behavioral expressions at the subconscious level, researchers aim to improve the accuracy and reliability of emotion classification at home and abroad. Multimodal emotion recognition methods often combine external behavioral expressions such as eye tracking and facial expressions to capture a more nuanced understanding of an individual's emotional state. While external behaviors such as facial expressions can provide intuitive emotional insights, they can sometimes lack authenticity and reliability. In contrast, physiological neural states provide a non-invasive and reliable means of assessing mood. By integrating techniques such as eye-tracking to analyze visual attention and cognitive processes, together with facial expression analysis, researchers can create a more robust framework for emotion recognition that considers both conscious and subconscious cues. This study lies in the potential of combining eye-tracking and facial expression analysis to improve emotion recognition. By exploring how these modalities complement each other in capturing emotional states, this study aims to contribute to the development of more accurate and reliable emotion recognition systems that can improve human-computer interaction and inform future developments in artificial intelligence.

We summarize our contributions as follows.

1. Based on relevant psychological theories, we design an emotion induction experiment and collect real-time eye movement signals and facial expression data. The experiment uses CAPS to prompt emotional responses (i.e., emotion activation).
2. We build a dual channel multimodal emotion recognition model using a CNN to extract and fuse spatiotemporal features of eye movements and facial expressions and to classify accurately three different emotions (positive, neutral, negative). Compared with single-mode emotion recognition, the fusion of eye movement signals and facial features has higher accuracy and reliability when identifying emotional states.

We structure our paper as follows. Section 2 introduces the related literature. Section 3 describes the production design and collection plan of the datasets. Section 4 discusses the design of the system model and the analysis of experimental results. Section 5 presents our conclusions.

2 Related Work

2.1 Emotion recognition based on facial expressions

As the external indicator of human emotions, facial expressions have always been one of the important research directions in the field of computer vision. Recent developments in facial emotion classification have made significant progress, with the shift from traditional to deep learning methods becoming significant. Traditional methods rely on manually designed features or shallow learning techniques, such as local binary patterns (LBPs) [4], local binary patterns from three orthogonal planes (LBP top) [5], non-negative matrix factorization (NMF) [6], and support vector machines (SVMs) [7]. However, these methods have difficulties recognizing emotions in actual environments. Improvements in computing performance have led facial emotion recognition to transition to deep learning. Of these, convolutional neural networks (CNNs) [8], VGGNet [9], residual neural network (RESNET) [10], and other deep learning methods have become the main research methods.

However, using facial expressions as the main basis for emotion recognition has many drawbacks. Different groups of people have varying degrees of concealment of facial expressions that make it difficult to understand the true emotions of the target audience. Emotion analyses are divided into two types: discrete models and dimensional models [11], with different researchers using different emotional quantification models. The emotion quantification model used in this article classifies emotions as positive, negative, or neutral.

2.2 Emotional recognition based on eye movement

Emotions can also be recognized using biological signals, facial expressions, speech intonation, and textual features. Recent developments in technology have led to collected eye movement signals as one of the classification features for emotion recognition. W. -L. Zheng et al [12] uses eye-tracking devices and cameras to record relevant information about research subjects while watching massive open online course (MOOC) videos and then classifies their learned emotions. S. Hickson et al [13] uses VR devices to record subject eye movement information and classifies the collected data into emotions. The accuracy of the method in five emotion categories has been effectively verified.

These results indicate that using eye movement signals for recognizing emotions has a certain degree of reliability. Currently, most researchers utilize eye movement features such as fixation, blinks, scans, and pupil data directly. We analyze the feature significance via deep learning and find that pupil diameter and gaze events are the main indicators of emotional state. We select pupil diameter (maximum, average, minimum), fixation time, number of fixation points, and first fixation time as classification features to capture emotional states.

2.3 Based on multimodal emotion recognition

More recently, research has focused on extracting features of different patterns, such as speech [14], expression [15], text [16], and physiological signals [17], and using machine learning methods to classify emotions. Emotion is expressed through multiple modes, and the use of a multimodal data fusion strategy for emotion classification has garnered significant research interest. At present, multimodal fusion uses speech and text [18], face and voice [19], EEG and voice [20], or electroencephalography (EEG) and face [21] for recognition.

Most multimodal emotion recognition uses a variety of external behavior performance modes, physiological neural states, and behavioral subconscious behavior. External performance behavior intuitively and effectively reflects individual emotions, but it lacks authenticity and reliability. Although the physiological nerve state is non-invasive and reliable, it is also unstable and has features that are difficult to select. At present, methods that combine external performance behavior and physiological neurological states are not currently widely used in emotion recognition research. Our approach uses expression and eye movements for modal fusion, which effectively avoids the difficulties of externally represented behaviors and ensures the stability of eye movement information reflecting emotions.

3 Data Management Pipeline

3.1 Selection of datasets

In the field of multimodal emotion recognition, existing datasets often fall short in capturing the diverse array of features necessary for comprehensive analysis. Many datasets primarily focus on facial expressions, neglecting other modalities such as eye movements, which are crucial for a holistic understanding of human emotions. Furthermore, the quality and accuracy of annotations in existing datasets may be compromised due to various factors such as data collection conditions and subjective biases of annotators, potentially leading to noise and inaccuracies that could impact model training and evaluation. To address these limitations, we conducted the development of a new dataset tailored specifically to the requirements of our research. This dataset collection process ensured both the quality of the data and the accuracy of annotations, providing a reliable foundation for our study. Moreover, our research focuses on emotion recognition within collaborative learning scenarios, where specific emotional stimuli may influence task performance. By designing our dataset collection protocol around these contextual factors, we aimed to ensure the relevance and applicability of the data to our research objectives. Through the creation of this new dataset, we not only addressed the gaps present in existing datasets but also contributed to the advancement of the field by providing a valuable resource for future research endeavors.

3.2 Datasets acquisition scheme

To study the characteristics of eye movements with changes in emotional states, it is necessary to obtain eye tracking data from subjects under different emotional states. We designed emotion induction experiments to stimulate emotional responses. In line with our categories, we induced three different types of emotions in our subjects: positive, neutral, and negative. Using materials from the International Affective Picture System (IAPS) and the Chinese Emotional Material Emotional Image System (CAPS), we selected a total of 135 images, dividing them into three groups of experiments. Each group of subjects underwent three rounds of positive, neutral, and negative experiments. We used an EyeLink1000Plus to record eye movement data and Lenovo high-definition cameras to record facial expression data. We selected a total of 16 subjects for the experiment. As subjects viewed the image materials, we obtained facial expression images through high-definition cameras [22] and recorded their eye movement information using an eye-tracking device. Our 16 participants were all college students with normal or corrected vision and consisted of 7 males and 9 females. To eliminate the influence of video viewing order, we used alternating playback between AB and BA to make the participants watch videos in different orders [23].

The experimental process is shown in Fig. 1. First, we explained the experimental content to the subjects and provided them with reading guidelines. In this experiment, we used the right eye alone because the movement of both eyes is conjugate, and the fixation positions of the two eyes are always very close. Therefore, apart from some paradigms of binocular information confrontation, there is no need to collect binocular information. Our left and right eyes differentiate into one dominant eye and one non-dominant eye, and theoretically collecting data from the dominant eye would be more accurate. However, in terms of practical operation, the determination of the dominant eye itself is itself controversial. Because more people are right-eye dominant, we chose that eye for all subjects. Due to the long duration of the experiment, a head fixator was used to restrict the movement of the subjects' heads in this experiment. We then calibrated of the eye tracker using the 9-point calibration method and a 1000 Hz sample rate. After the calibration, we began the experiment, with the subjects watching the carousel images displayed on the computer screen. After the image was displayed for 4 seconds, a calibration eye screen was shown. To prevent the subject's pupils from disappearing, a second image appeared, with the whole sequence performed 15 times. During the process of rotating images, as shown in Fig. 2, we collected data regarding the subject's facial expressions and eye movements through high-definition cameras and eye-tracking devices.

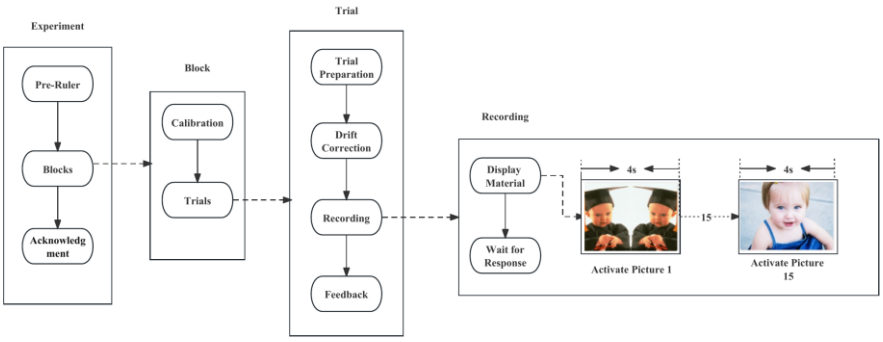

Fig. 1. Flow chart of the eye movement signal collection experiment.

In this data collection experiment, we collected the data under uniform indoor conditions, with a face video sampling rate of 30 Hz and an EyeLink1000Plus eye tracker sampling rate of 1000 Hz. The total duration of collected valid data exceeded 200 minutes. Due to improper experimental operation, some eye movement data of three subjects were lost, possibly due to prolonged experimental time when the subjects moved their heads, and their pupils were not captured by the eye tracking device. After data cleaning, the effective rate of eye tracking data exceeded 85%. The number of events reported by the retained subjects was 883 positive events, 1018 negative events, and 634 neutral events. Each event only represents one emotional state. The eye movement data collected in the experiment includes gaze data, scanning data, and pupil data.

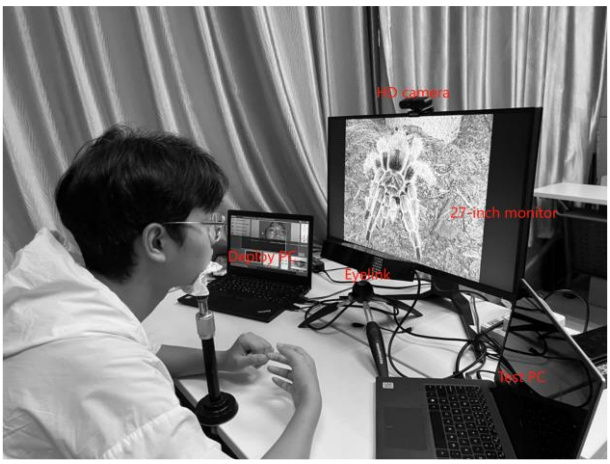

Fig. 2. Data collection experimental environment.

3.3 Data preprocessing

The facial data initially collected in the experiment is continuous, making it difficult to align the continuous facial image features with the discrete eye movement data features in the model. Thus, frame segmentation was required. We labeled the collected video information and used Python's CV2 module VideoCapture to read the frames of the original video. We set the time interval for selecting frames to 6 seconds, with each round of images displayed for 4 seconds. Assuming a Calibration time of 2 seconds, a set of data was taken every 6 seconds when changing events. Fig. 3 shows a sample of our data, where the image was first grayscale, and then enhanced in contrast and brightness via a histogram equalization. Then, a Gaussian filter was used to smooth the image to reduce the effect of noise. Finally, the image size was standardized to a uniform size to ensure the consistency of the subsequent facial feature extraction and analysis process. The Viola–Jones facial detection algorithm was used to identify key features in facial images, such as eyes, nose, and mouth. The face key point detector in the Dlib library was used to extract the coordinate information of facial landmarks, with the feature vectors of facial expressions calculated based on these feature points, including the intensity and direction of facial expressions.

Fig. 3. Sample facial expression images.

The original eye movement data collected in the experiment was divided into fixation, scanning, and event classification. We used the eye tracking instrument and experimental visualization software DataViewer to divide the interest area and extract the event data within the interest area. Due to the extracted data being stored in a CSV file, abnormal data needs to be filtered. We removed fixation points less than 100 ms and or greater than 1000 ms. Due to the high susceptibility of pupil diameter to light exposure, we used principal component analysis (PCA) to remove the influence of the first principal component (light).

Suppose Y is the $M \times N$ matrix representing pupil diameters to the same video clip from N subjects and M samples. Then $Y = A + B + C$, where A is luminance influences which is prominent, and B is emotional influences which we want, and C is the noises. We use principal component analysis to decompose Y. We extract the first principal component from PCA to approximate the pupil response for the lighting changes during the experiments.

4 Model

Based on the basic dataset obtained from the preceding experiment, we designed a multimodal emotion recognition model for facial expressions and eye movement signals using a CNN. This model combines feature-level fusion and the VGG Inspired LightNet structure to perform emotion recognition using eye movement data and image data. Deep learning models automatically learn complex features from input data and generate an effective classification of emotions through an end-to-end training process.

4.1 Multimodal model architecture

High-quality facial expression data and eye movement signal data are key to emotion classification and feature saliency analysis. Therefore, after each portion of the experiment, we have the participants complete the Self Positive and Negative Emotion Scale (PANAS) to determine the reliability of emotion induction.

We use a random forest to evaluate the importance of eye movement features and ultimately select pupil diameter (maximum, average, minimum), fixation time, number of fixation points, and first fixation time as classification features. Preprocessing and feature extraction have been optimized for classification to reduce training time and noise influence. Event-based concatenation is performed on the two types of data to better fit the input of the model. Finally, different machine learning algorithms are used to construct a single modal classification model, which is then trained and evaluated.

When processing the two types of data (eye movement signals and facial expressions), the model first learns the spatial image features using a CNN and then learns their temporal features using a Long Short-Term Memory (LSTM) network. In this way, the model can efficiently capture the temporal information of both data modalities. We used the Keras framework to build the model, as shown in Fig. 4. The model is divided into two channels for input, with eye movement and facial expression data trained through the two models. Eye movement signals are generated through real-time tracking, while expression modalities are transmitted through static images. Therefore, we adopted two feature extraction methods. Convolutional neural networks were used to learn the spatial image features of state frames, and then LSTM was used to learn their temporal features. The LSTM can remember and use information from previous time steps, leading to a better understanding of temporal dependencies between data, which helps to identify temporal patterns and changes in emotional states. We used the VGG Inspired LightNet model for convolutional neural networks. Inspired by VGGNet [24], we used fewer convolution kernels and shallower convolution layers. Adopting a stacked structure of convolutional and pooling layers, features were extracted from images using different convolutional kernels. A Dropout layer was introduced after each convolutional block in the convolutional neural network to randomly discard a portion of neurons during training to prevent overfitting. Using the concatenate layer to fuse the features of eye movement data and image data, we create a representation that integrates multimodal information. Finally, these extracted features will be used as inputs to XGBoost and these combined features are used to predict the emotional state (positive, neutral or negative) through the XGBoost layer. XGBoost is a boosted-tree model

that is efficient, flexible, generalizable, feature-engineering friendly, and has good interpretability when dealing with small datasets.

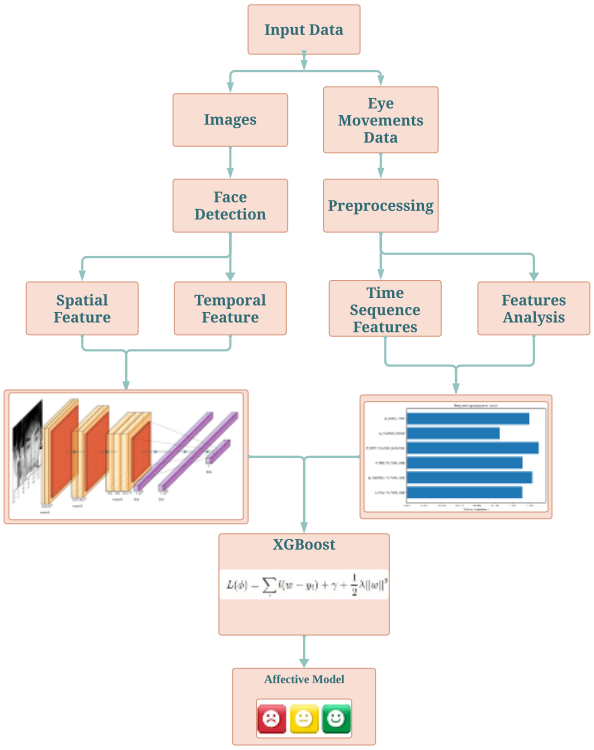

Fig. 4. The network structure of dual channel emotion recognition for eye-tracking expressions.

4.2 Feature fusion

The role of eye movement signals in facial expressions is to provide additional information and cues to help describe an individual's emotional experience more fully. By combining eye movement signals and facial expressions, the model can identify emotional states more accurately, improving the accuracy and reliability of identification. Eye movement signals can reflect an individual's visual attention and cognitive processes, while facial expressions can convey emotional states, and the combination of the two can provide richer information for emotion recognition.

Feature fusion mainly includes feature layer fusion, classification decision layer fusion, and collaborative computing methods. Feature layer fusion focuses on exploring the feature fusion of multimodal data with different granularities, enabling the organic combination of different granularities of facial expressions and eye movement modalities, and fully considering the temporal nature of emotions while retaining the importance of global features in emotion recognition. For the fusion method of multi-

modal classification decision layers, by comparing the performance of different classifiers in single-modal emotion classification of facial expressions or eye movements, as well as bimodal emotion classification of facial expressions and eye movements, the influence of individual differences and emotion categories on classifier selection is deeply studied by the system. Combined with the research on different classifiers expressing eye movement patterns, we propose an adaptive adjustment of the weight of facial expressions and eye movement bimodal. We integrate the classification results through a voting mechanism. The multimodal collaborative computing method emphasizes the collaborative use of feature-level fusion and decision-level fusion methods. Different classification calculation modes are synergistically applied and fused at the feature layer for different granularities and frequencies of expression and eye movement patterns.

In feature-level fusion, feature vectors from different methods are concatenated to form a larger feature vector. In our experiments, we select differential entropy features from eye movement data and facial expression responses and train a fusion model that combines eye movement features and facial expression features. For decision-level fusion, the two classifiers are trained separately using different features and fused using certain principles or learning algorithms to generate new classifications. We applied two principles of decision-level fusion in our study. One is the maximum strategy, which selects the high probability output of a classifier trained with a single modality alone as the result. The other is a summation strategy, which summarizes the probabilities of the same sentiment from different bands and selects the higher one.

In the experimental scenario using eye movement data, due to the temporal synchronization between facial data and eye movement information, the two modalities are triggered based on the same event. Fig. 5 shows the process of multimodal feature fusion by extracting spatiotemporal features of eye movements and expressions. Event-based concatenation not only better captures contextual information but also facilitates the model in capturing temporal relationships between events.

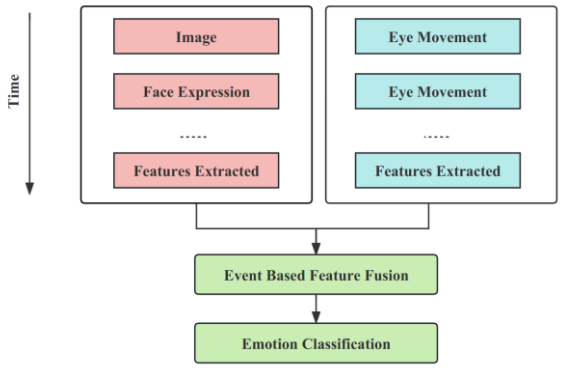

Fig. 5. Feature fusion strategy.

5 Results

This section presents our dataset, the significance analysis of data features, parameter settings, and experimental results.

5.1 Generation of datasets

We displayed the selected stimulus material on a 27-inch high-definition display screen and replaced the image every 4 seconds. Simultaneously, we used a Lenovo Thinkplus WL24A high-definition camera to record the facial expressions of the subjects and an Eyelink1000plus to record the eye movement data. The data collection was conducted under indoor conditions with uniform lighting, with a face video sampling rate of 30 Hz and an eye tracking device sampling rate of 1000 Hz.

The eye movement data is collected as a long string of continuous values. Due to the low frequency of pupil behavior, we adopt a sub-sampling method for each trial. Due to the calibration time t_c between each round, we denote the image presentation time for each round as t_m and the first sampled data as t_0 and define the relationship between the sub-sampling window time U_t and the trial i as:

$$U_t(i) = t_0 + (i-1) * t_c + (i-1) * t_m \tag{1}$$

The total number of samples in the datasets was 15210 (divided into 2535 samples by topic).

5.2 Feature significance analysis

We conducted feature significance analysis on features and emotional labels using the collected data. As shown in Fig. 6, we used a box plot to perform statistical analysis on each feature of the sample, displaying the distribution of features under different emotional categories. The 0 in the horizontal axis indicates a negative emotion, 1 indicates neutral, and 2 indicates positive.

We analyzed three types of pupil size characteristics: maximum, average, and minimum. For the maximum pupil size, we observed that there were multiple outliers in the data under all three labels, with a very wide range of values. Compared with labels 1 and 2, the median of label 0 was slightly lower, and the interquartile range (IQR) of label 0 was also narrower than other labels. For the average pupil size feature, we found that the median of label 0 was lower, and the IQR was more compact than labels 1 and 2, while labels 1 and 2 were similar. All labels had outliers, with the extreme outlier of label 2 being particularly significant. Finally, regarding the minimum pupil size feature, we observed that the median of label 0 was significantly lower than labels 1 and 2, and label 0 had many lower outliers.

Overall, we observed significant differences in pupil size and gaze characteristics with different experimental conditions or groups. Specifically, label 0 showed lower values in all pupil size and gaze features, and the data distribution was more compact, while labels 1 and 2 exhibited similar or different patterns. These findings may reflect

differences in visual attention and cognitive processing under different conditions, providing important clues for further understanding eye movement behavior.

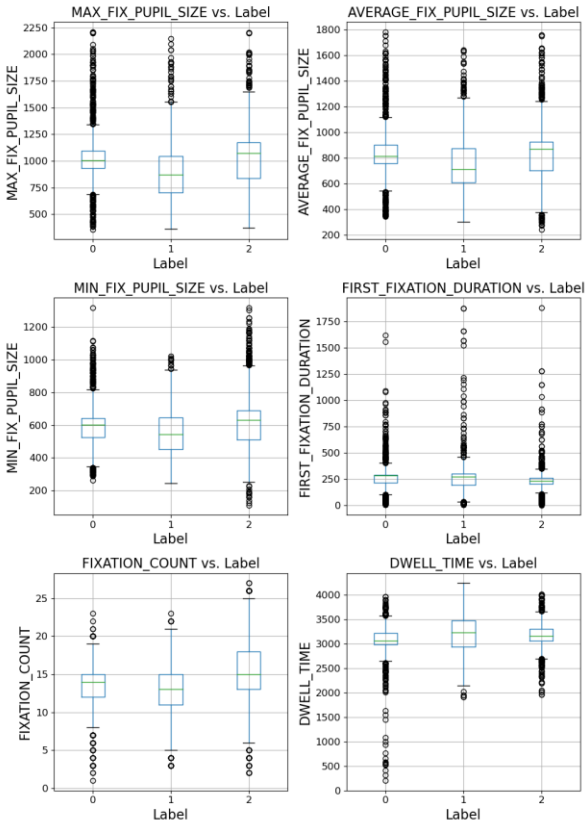

Fig. 6. Analysis of average pupil size in a certain trial with different subjects.

5.3 Experimental parameters and results

This experiment was run on a server with a CPU of i5-13600kf, a GPU of RTX4070ti, and a running memory of 32GB. The experimental setup is mainly based on Python language, with model training set to 200 rounds and a batch size is 64.

Fig. 7 shows the accuracy of different modalities in emotion recognition. In the multimodal versus unimodal experiments, a random forest algorithm was used to predict eye movement data, and a CNN was used for facial expression training, with VGG Inspired LightNet used for multimodal training. The data shows that the accuracy of the multimodal method in negative, neutral, and positive emotions was 98%, 93%, and 96%, respectively, which is significantly better than the single modal method ("eye movement" and "facial expression"). This shows that, compared with a single mode, combining multiple modes is more effective for emotion recognition.

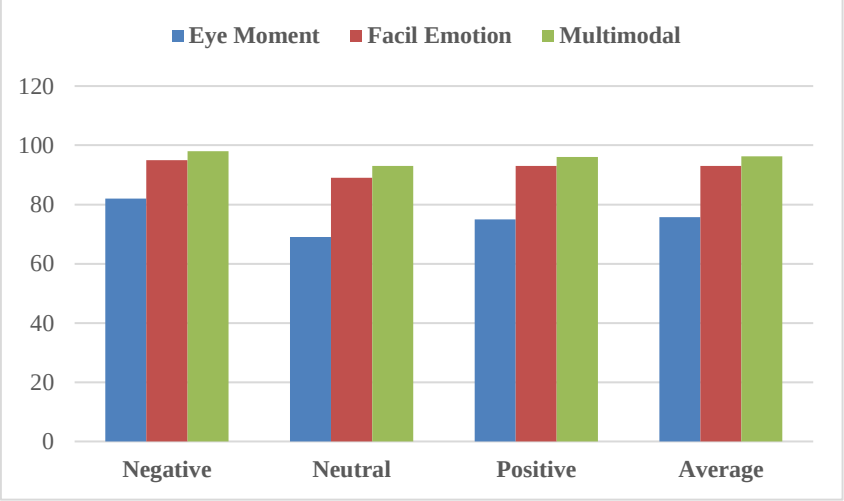

Fig. 7. Analysis of average pupil size in a certain trial with different subjects.

6 CONCLUSION

In this article, we present our construction of an effective emotional activation design scheme using psychological emotional theory and employing eye-tracking devices and cameras to record eye movement signals and facial expressions. Based on these data, we perform event-based concatenation and adopt feature layer fusion. We also design a dual-channel multimodal emotion recognition model using a CNN that effectively captures the temporal information of two modal data. In experiments, this model achieved an average accuracy of 96.25% when classifying emotions. In future work, we will attempt to increase the fusion of multiple modalities by incorporating electroencephalography (EEG).

Acknowledgment. Thanks to Beijing Borun Vision Technology Co., Ltd., for generously providing the eye tracking equipment Eyelink1000Plus, as well as the synchronization platform Experiment Builder and DataViewer, along with their invaluable technical guidance. We are also grateful to all the experimenters and subjects who participated in the data collection process.

This work was supported in part by the National Natural Science Foundation of China under grant 62207033, and the Teaching and Research Projects of National Ethnic Affairs Commission of the People's Republic of China, under grant of 23092.

References

1. Z. Zeng, M. Pantic, G. I. Roisman and T. S. Huang. : A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39-58, Jan. 2009.

2. Pan, B., Hirota, K., Jia, Z., and Dai, Y.: A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. Neurocomputing, 561, 126866. 2023.
3. Ahmed, N., Al Aghbari, Z., and Girija, S.: A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent Systems with Applications, 17, 200171. 2023.
4. J. K. Josephine Julina and T. S. Sharmila.: Facial Emotion Recognition in Videos using HOG and LBP. 2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), Bangalore, India, pp. 56-60, 2019.
5. Y. Wang, Hui Yu, B. Stevens and Honghai Liu.: Dynamic facial expression recognition using local patch and LBP-TOP. 2015 8th International Conference on Human System Interaction (HSI), Warsaw, Poland, pp. 362-367, 2015.
6. J. Li and M. Oussalah.: Automatic face emotion recognition system. 2010 IEEE 9th International Conference on Cyberntic Intelligent Systems, Reading, UK, pp. 1-6, 2010.
7. M. H. Abdul-Hadi and J. Waleed.: Human Speech and Facial Emotion Recognition Technique Using SVM. 2020 International Conference on Computer Science and Software Engineering (CSASE), Duhok, Iraq, pp. 191-196, 2020.
8. S. Begaj, A. O. Topal and M. Ali.: Emotion Recognition Based on Facial Expressions Using Convolutional Neural Network (CNN). 2020 International Conference on Computing, Networking, Telecommunications & Engineering Sciences Applications (CoNTESA), Tirana, Albania, pp. 58-63, 2020.
9. O. Khajuria, R. Kumar and M. Gupta.: Facial Emotion Recognition using CNN and VGG-16. 2023 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, pp. 472-477, 2023.
10. X. Shen, X. Xu and Y. Zhuang.: Facial Emotion Recognition Based on Sobel-Resnet. 2021 IEEE 5th Information Technology,Networking,Electronic and Automation Control Conference (ITNEC), Xi'an, China, pp. 484-488, 2021.
11. Yue Z, Lofi C, Hauff C.: Scalable Mind-Wandering Detection for MOOCs: A Webcam-Based Approach. European Conference on Technology Enhanced Learning. Springer, Cham. Tallinn, Estonia. pp: 330-344. 2017.
12. W. -L. Zheng, B. -N. Dong and B. -L. Lu.: Multimodal emotion recognition using EEG and eye tracking data. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, pp. 5040-5043, 2014.
13. S. Hickson, N. Dufour, A. Sud, V. Kwatra and I. Essa.: Eyemotion: Classifying Facial Expressions in VR Using Eye-Tracking Cameras. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, pp, 2019.
14. R. Sato, R. Sasaki, N. Suga and T. Furukawa.: Creation and Analysis of Emotional Speech Database for Multiple Emotions Recognition. 2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Yangon, Myanmar, pp. 33-37, 2020.
15. A. Kartali, M. Roglić, M. Barjaktarović, M. Đurić-Jovičić and M. M. Janković.: Real-time Algorithms for Facial Emotion Recognition: A Comparison of Different Approaches. 2018 14th Symposium on Neural Networks and Applications (NEUREL), Belgrade, Serbia, pp. 1-4, 2018.
16. M. -H. Su, C. -H. Wu, K. -Y. Huang and Q. -B. Hong.: LSTM-based Text Emotion Recognition Using Semantic and Emotional Word Vectors. 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), Beijing, China, pp. 1-6, 2018.
17. S. M. Alarcão and M. J. Fonseca.: Emotions Recognition Using EEG Signals: A Survey. IEEE Transactions on Affective Computing, vol. 10, no. 3, pp. 374-393, 1 July-Sept. 2019.

18. U. A. Asiya and V. K. Kiran.: A Novel Multimodal Speech Emotion Recognition System. 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, pp. 327-332, 2022.
19. L. Cai, J. Dong and M. Wei.: Multi-Modal Emotion Recognition from Speech and Facial Expression Based on Deep Learning. 2020 Chinese Automation Congress (CAC), Shanghai, China, pp. 5726-5729, 2020.
20. D. S. Moschona.: An Affective Service based on Multi-Modal Emotion Recognition, using EEG enabled Emotion Tracking and Speech Emotion Recognition. 2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), Seoul, Korea (South), pp. 1-3, 2020.
21. Y. Chen, Z. Bai, M. Cheng, Y. Liu, X. Zhao and Y. Song.: Multimodal Emotion Recognition for Hearing-impaired Subjects by Fusing EEG Signals and Facial Expressions. 2023 42nd Chinese Control Conference (CCC), Tianjin, China, pp. 1-6, 2023.
22. S. P. K. Malladi, J. Mukherjee, M. -C. Larabi and S. Chaudhury.: EG-SNIK: A Free Viewing Egocentric Gaze Dataset and Its Applications. IEEE Access, vol. 10, pp. 129626-129641, 2022.
23. S. Droit-Volet, S. Fayolle, and S. Gil. Emotion and time perception: Effects of film-induced mood. Frontiers in Integrative Neuroscience, vol. 5, Art. no. 33, 2011.
24. K. Simonyan and A. Zisserman.: Very deep convolutional networks for large-scale image recognition. 2015 3rd International Conference on Learning Representations (ICLR 2015), pp. 1-14, 2015.