Neural Computing and Applications (2023) 35:24713–24725
https://doi.org/10.1007/s00521-023-08366-7

S.I.: APPLICATIONS AND TECHNIQUES IN CYBER INTELLIGENCE (ATCI2022)

Multimodal sentiment system and method based on CRNN-SVM

Yuxia Zhao[1,2,3] · Mahpirat Mamat[1,4] · Alimjan Aysa[1,4] · Kurban Ubul[1,4]

Received: 30 August 2022 / Accepted: 13 February 2023 / Published online: 11 March 2023

Abstract
Traditional sentiment analysis focuses on text-level sentiment mining, transforming sentiment mining into classification or regression problems, resulting in a sentiment analysis low accuracy rate. Sentiment analysis refers to the use of natural language processing, text analysis, and computational linguistics to systematically identify, extract, quantify, and study sentimental states. Therefore, more scholars have begun to focus on speech recognition and facial expression recognition research, and extracting and analysing people's sentiment tendencies can improve sentiment recognition accuracy. Traditional single-modal sentiment analysis can no longer meet people's needs. Therefore, this paper proposes a multimodal sentiment analysis method based on the multimodal sentiment analysis method that can obtain more sentimental information sources and help people make better decisions. The experimental results in this paper show that the highest recognition rates of CNN-SVM, RNN-SVM, and CRNN-SVM were 76.8%, 71.2%, and 93.5%, respectively. It can be seen that CRNN-SVM has the highest sentiment tendency recognition rate in deep learning, so it is suitable to apply CRNN-SVM to sentiment tendency analysis system design in this paper. The average accuracy rate of the system designed in this paper was 91%, and the stability was also very strong, which shows that the system designed in this paper is meaningful. The main contribution of this paper is based on the limitations of single-mode emotion analysis. It proposes a multimode emotion analysis method and introduces a convolutional neural network to help people obtain more emotional information sources to meet their needs.

Keywords Sentiment analysis · Deep learning · Convolutional neural network · Multimodal approach

1 Introduction

In daily communication, humans can recognize each other's emotional changes by listening to language and observing expressions and gestures, recognizing emotional state information, and then conducting emotional communication. However, for a machine to be able to perceive and understand emotions like human beings, the machine must be enabled to simulate human capabilities in this regard so that the machine can capture multimodal emotional features, process them, and finally express the corresponding human emotional capabilities. In recent years, with the rapid development of the internet, an increasing number of people have published information on the internet, which makes the internet a very large information warehouse. Due to the continuous expansion of internet information, it is increasingly difficult for people to find the information they want on the internet. The emergence of search engines has decreased the time for people to obtain information to a certain extent, but with the diversified development of

✉ Kurban Ubul
kurbanu@xju.edu.cn

Yuxia Zhao
zyx@stu.xju.edu.cn

Mahpirat Mamat
xmahpu76@163.com

Alimjan Aysa
alim@xju.edu.cn

[1] School of Information Science and Engineering, Xinjiang University, Ürümqi 830046, Xinjiang, China

[2] School of Mathematics and Computer Applications, Shangluo University, Shangluo 726000, Shaanxi, China

[3] Engineering Research Center of Qinling Health Welfare Big Data, Universities of Shaanxi Province, Shangluo 726000, Shaanxi, China

[4] Xinjiang Laboratory of Multi-Language Information Technology, Ürümqi 830046, Xinjiang, China

information on the internet, people have more detailed and personalized requirements for information acquisition. Sentimental inclination has become a personalized indicator for people to filter information. Sentimental inclination is the tendency of the subject's inner likes and dislikes to the subjective existence of a certain object and the inner evaluation.

This paper takes multimodality and RAMAS as the research object and conducts research on multimodal perception verification of speech, voice, and signal. For the speech module, this paper introduces a new CNN-spatial–temporal LSTM deep tissue extraction method and, on this basis, introduces the spatial–temporal inclusion focus. This study combines traditional spatiotemporal highlighting operations with convolutional brain network design to strengthen the extraction of high-attention components in information representation groups and uses different LSTM structures to determine the spatial and transient correlations between elements. Finally, features showing modularity are obtained. This paper discusses the perceptual cognitive effect of single-module speech under different time and space block boundary conditions and discusses the effect of the multimodule combination method for speech module recognition. Information is obtained in voice mode, and the sound effects in emobase2010 are removed through the openSMILE program.

The data were processed to obtain the required samples. Using the Dlib facial recognition library and the OpenCV computer vision library, the raw images in the database were extracted and analysed in the form of several consecutive facial expression images. Using the SciPY library to edit the original video containing special emotional components, the speech pattern of this part was obtained. Kinect's skeleton data were selected from the library, and a script to obtain the pose mode data was written.

The concept of deep learning is introduced. Based on this, this paper proposes a method that uses traditional space–time feature key points to improve the attention weight of a convolutional neural network, extracts the features of expression modality data from continuous images, and compares them in the application of expression unimodal emotion recognition.

We briefly introduce the basic canonical correlation analysis principle and propose two canonical correlation analysis methods on this basis.

Using the above feature extraction method, we extract the expression mode features in the two databases, use openSMILE to extract the corresponding voice mode features, use the selected gesture mode to conduct multimode fusion experiments, and use PCA in simple series fusion, multiclass canonical correlation analysis, multiclass kernel canonical correlation score, decision-level fusion and other fusion techniques are analysed and compared. From the two perspectives of emotional features and emotional feature fusion, facial expression patterns based on improved convolutional neural networks are studied, combined with supervised least squares multiclass kernel canonical correlation analysis and sparse supervised least squares multiclass A method for nuclear canonical correlation analysis.

Sentiment analysis is also called opinion mining, and its purpose is to mine comment objects and opinions about the object from documents or document collections. This technique often employs closely related information extraction techniques to discover objects in text and their corresponding viewpoints. As a cross-field research hotspot, text sentiment analysis is involved in many fields, such as natural language processing, information retrieval, automatic summarization, and data mining. Therefore, sentiment analysis technology has broad application prospects in the user comment analysis and decision-making field. The innovation of this paper is that it proposes a multimodal sentiment analysis method based on AI deep learning and applies it to the system designed in this paper to improve the system's ability to analyse sentiment. Different information modes need different processing and modelling methods. The core drive of the multimodal method is that more information sources can help people make better decisions. A multimodal model strategy is necessary for emotion analysis tasks. First, in many cases, it is difficult to accurately judge the emotional state only by text or voice. An extreme example is irony. Irony often combines neutral or positive text content and an audio expression that does not match the content to complete a negative (negative) emotional expression. This kind of situation is difficult to solve fundamentally only by a single mode. Second, the single-mode model is easily affected by noise, resulting in effect problems.

2 Related work

Natural language processing is an important direction in the computer science and artificial intelligence fields. It studies various theories and methods that can realize effective communication between humans and computers using natural language. Sentiment analysis belongs to the category of natural language processing and data mining, which is closely related to information retrieval, statistical analysis, and other techniques. Sentiment analysis techniques have a wide range of uses in real life. Young IF examined specific relationships between various environmental experiences and sentimental dispositions. He combined the theories of geographers and psychologists. He showed that the fusion and separation of people and the environment could be manipulated through experiments to have causal effects on sentiment experience [1]. Yuan Z

proposed microexpression recognition to infer the true sentiments that people are trying to hide from video clips of faces. This is a very challenging task because microexpressions are of low intensity and short duration, which makes microexpressions difficult to observe [2]. Bertsch K found that difficulty controlling sentimental impulses is an important component of borderline personality disorder (BPD), often leading to destructive, impulsive behaviour towards others [3]. Yan J believed that sentiment orientation analysis was a key issue in endowing artificial machines with true intelligence in many large-scale potential applications. Electroencephalogram (EEG) signals and video face signals are widely used to track and analyse human sentiment information as external representations of human sentiment [4]. Alghifari M F proposed the challenging nature and broad future prospects of speech sentiment analysis (SER). He utilized deep neural networks (DNNs) to identify human speech sentiments, using optimized networks to introduce and validate a custom database [5]. Scholars have suggested that sentiment analysis has broad prospects for development. It is also of great significance to people's daily work. If a highly accurate sentiment analysis system can be constructed, it will make greater progress in sentiment analysis. Academics also did not describe how to build the system.

AI deep learning has played an important role in multimodal sentiment analysis. Mittal T proposed a method based on convolutional neural network learning for multimodal sentiment recognition. His method combines prompts from multiple common modes (such as face, text, and voice) and is more powerful than other methods. It can detect noise in any single mode [6]. Huan R H proposed a multimodal sentiment recognition method based on a convolutional neural network to improve sentiment recognition accuracy in the time context. A new network initialization method is proposed and applied to the convolutional neural network model, which can further improve sentiment recognition accuracy. This method can improve sentiment recognition accuracy in three single modes of text, vision, and audio and improve multimodal sentiment recognition accuracy in the video. This method is superior to existing multimodal sentiment recognition methods in sentiment classification and sentiment regression [7]. Deep learning in sentiment analysis can improve the accuracy of analysis and a high rate of sentiment analysis, which enables people to better analyse sentimental tendencies and better understand people's sentiments to solve problems. Therefore, it is very meaningful to apply AI deep learning to multimodal sentiment analysis systems.

3 Sentiment tendency method based on deep learning

3.1 Convolutional neural network (CNN) sentiment tendency model

Sentiment analysis technology is the most frequently used in the field of user comment analysis and decision-making, and it has the most commercial value [8, 9]. In addition, sentiment analysis is also used in public opinion monitoring, information prediction, question-answering systems, and other fields. Judging the sentiment tendency of articles on the internet can provide decision support for individuals, governments, and enterprises.

A convolutional neural network is a feedforward neural network that is well applied to the analysis of visual images [10]. The difference between a CNN and an ordinary neural network is as follows. First, the local connection is used instead of the full connection, which greatly reduces the parameters that need to be learned in the network. Second, multiple filters are set in each layer, and each filter can extract different sample features. The third is to set the downsampling layer, which further reduces the number of parameters. The basic structure of the CNN is shown in Fig. 1:

Figure 1 shows the difference between nn and full connection: The original calculation of an output neuron is connected to all the neurons of the input layer, but now it is only connected to the neurons of the local input layer. The basic structure of a CNN includes convolutional layers and sampling layers. CNN can automatically extract multidimensional features from input samples through the design of filters, and through the local connection and sampling mechanism, the connections in the neural network are greatly simplified, which not only reduces the difficulty in manually extracting features. It also improves the training speed of the model. In practical research, CNNs are also increasingly applied to various fields [11].

The neural network can fit any function, and the basic neural network unit is the neuron, that is, the perceptron. The input of the perceptron becomes a layer of neurons, and the input of the neuron model is the output signal of other neurons. The basic method of this paper is to use the CBOW model in word2vec to pretrain word vectors and then input the word vectors to the CNN model, which is a fine-tuning and modification of the CNN model [12]. Word vectors are a collective term for a set of language modelling and feature learning techniques in embedded natural language processing, where words or phrases from a vocabulary are mapped to vectors of real numbers. Its calculation formula is Formula 1:

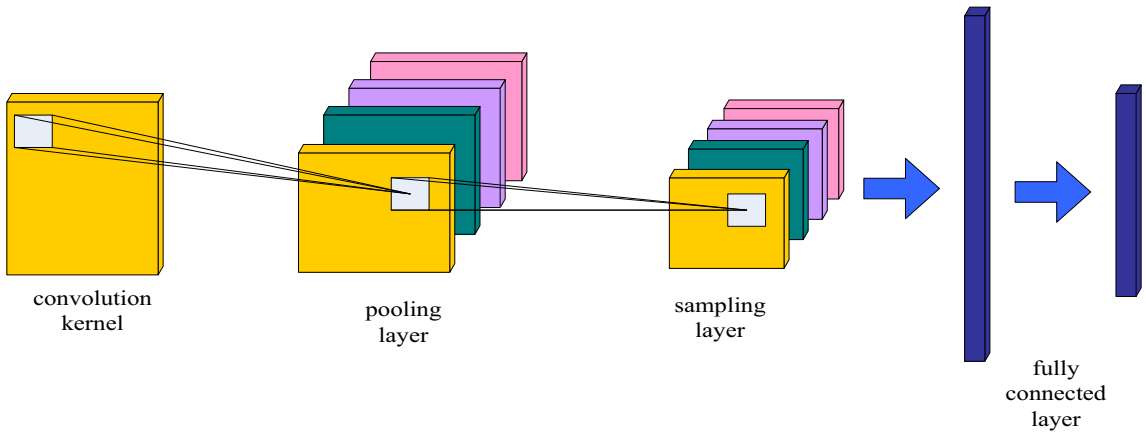


Fig. 1 Basic structure of CNN

The threshold, also called the critical value, refers to the lowest or highest value that an effect can produce. This term is widely used in various fields. The basic method in this paper is to use the CBOW model in word2vec to pretrain threshold for word vectors and then input the word vectors into the CNN model, which is a fine-tuning and modification of the CNN model [12]. Its calculation formula is Formula 1:

$$\text{output} = \begin{cases} 0, \sum_j w_j a_j \leq \text{threshold} \\ 1, \sum_j w_j a_j \succ \text{threshold} \end{cases} \tag{1}$$

The neuron makes some adjustments on the basis of the perceptron so that:

$$w^T A = \sum_j w_j a_j \tag{2}$$

Then, move the threshold to the other side of the inequality and name it bias, as in Formula 3:

$$b = -\text{threshold} \tag{3}$$

From this, the mathematical model of a single neuron can be written as Formula 4:

$$\text{output} = \begin{cases} 0, w^T A + b \leq 0 \\ 1, w^T A + b \succ 0 \end{cases} \tag{4}$$

The way information is propagated in a feedforward neural network is Formula 5:

$$z^{(l)} = W^{(l)} \cdot a^{(l-1)} + b^{(l)} \tag{5}$$

The most commonly used algorithm to minimize the loss function is "gradient descent". There are many methods to classify the loss function, which can be divided into the empirical risk loss function and structural risk loss function according to whether regular items are added. The purpose of this paper is to learn the weight W and its bias b of each sample by minimizing the loss function, which requires the use of the cross-entropy loss function as Formula 6:

$$L\left(b, \hat{b}\right) = -b^T \log \hat{b} \tag{6}$$

$\hat{b}$ is the sample label value generated during the calculation.

The goal of this paper is to minimize $\partial R(W, b)$ so that the network parameters can be learned by gradient descent. In each iteration, the parameter $W^{(l)}$ of the lth layer is updated as Formula 7:

$$W^{(l)} = W^{(l)} - \alpha \frac{\partial R(W, b)}{\partial W^{(l)}} \tag{7}$$

α is the learning rate. For learning parameters in the neural network, the backpropagation algorithm is usually used.

3.2 Recurrent neural (RNN) sentiment tendency model

In the RNN model, information is only propagated backwards between layers, and there is no information transmission between neurons in the same layer, so there is a problem: the ability of the feedforward neural network to process sequential data is poor [13]. A feedforward neural network is the simplest type of neural network. Each neuron is arranged in layers, and each neuron is only connected to the neurons in the previous layer, as shown in Fig. 2:

As shown in Fig. 2, a recurrent neural network is an artificial neural network with a tree-like hierarchical structure in which the network nodes recurse the input information in the order of their connection. Especially in natural language, assuming that each word is input as a feature, each word is not independent of the other but has a contextual relationship. A recurrent neural network is designed for this kind of sequential data [14]. The state

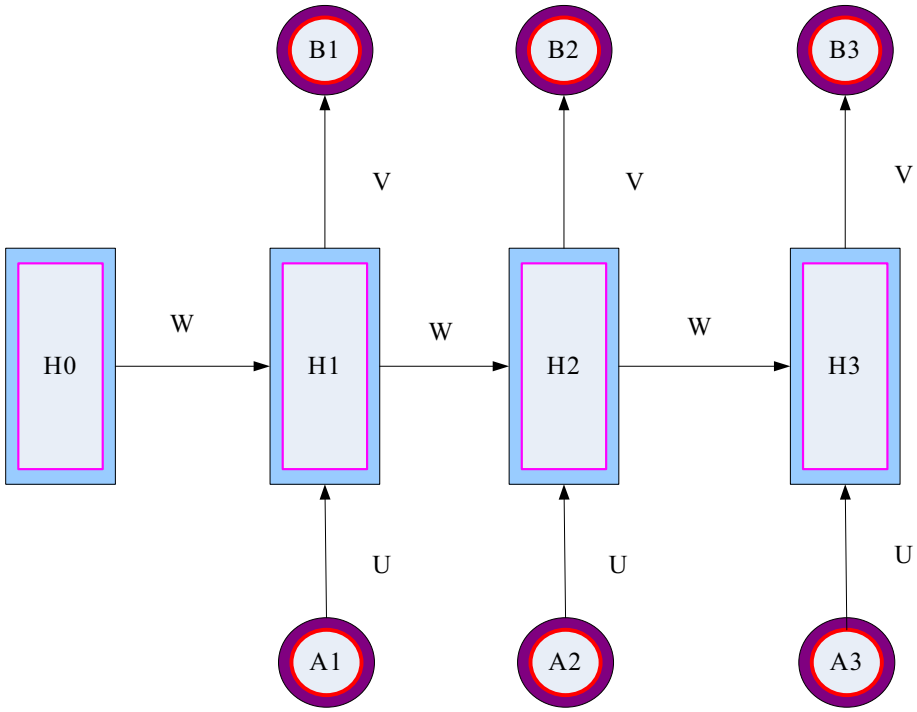


Fig. 2 Schematic diagram of a recurrent neural network

function of the hidden layer at the current moment is usually expressed by the following formula:

$$h_t = f(Uh_{t-1} + Wa_t + b) \tag{8}$$

where U is the state weight matrix, $f(\cdot)$ is a nonlinear activation function, and some unimportant information can be selectively forgotten.

In fact, the gate is a fully connected layer [15]. Assuming that at time t, the memory unit in the LSTM is denoted as c_t, and all historical information is controlled by three gates whose values are between [0, 1]. The input gate i_t is Formula 9:

$$i_t = \sigma(W_t a_t + U_i h_{t-1} + V_i c_{t-1}) \tag{9}$$

The forget gate is Formula 10:

$$f_t = \sigma\left(W_f a_t + U_f h_{t-1} + V_f c_{t-1}\right) \tag{10}$$

The output gate is Formula 11:

$$o_t = \sigma(W_o a_t + U_o h_{t-1} + V_o c_{t-1}) \tag{11}$$

Similar to CNN, RNN maps the preprocessed one-dimensional text into a two-dimensional vector. Unlike CNN, instead of inputting the entire document matrix at one time, the word vectors in the document are input into the feature extraction layer one by one in the order of words according to the time series [16]. At time t, the parameters in the LSTM are updated as in Formula 12:

$$h_t = o_t \oplus \tanh(c_t) \tag{12}$$

where c_t represents the input at the current moment and $\tanh(\cdot)$ is the activation function. Through this gate mechanism, LSTM can learn long-term historical information, such as Formula 13:

$$c_t = f_t \oplus c_{t-1} + i_t \oplus \tilde{c}_t \tag{13}$$

In the past few years, recurrent neural networks have shown better performance in natural language processing, especially various variants of LSTM.

3.3 Multimodal sentiment tendency based on deep learning

Multimodal sentiment analysis often involves the fusion of two modalities and generally includes two different fusion methods: feature fusion and decision fusion [17]. Feature fusion means that the features of two modalities are analysed and fused into a new feature, and a classifier is trained to classify it. Decision fusion refers to the analysis based on the results of each modal classification. This paper studies the above two different fusion methods and analyses the support vector machine (SVM) theory commonly used in multimodal sentiment analysis. The schematic diagram of SVM is shown in Fig. 3:

As shown in Fig. 3, SVM has developed rapidly and derived a series of improved and extended algorithms, which have been applied in pattern recognition problems such as portrait recognition and text classification. SVM is a kind of binary classifier. The greatest difference from the perceptron is that the classifier needs to meet the precondition of the largest interval, which makes the classifier

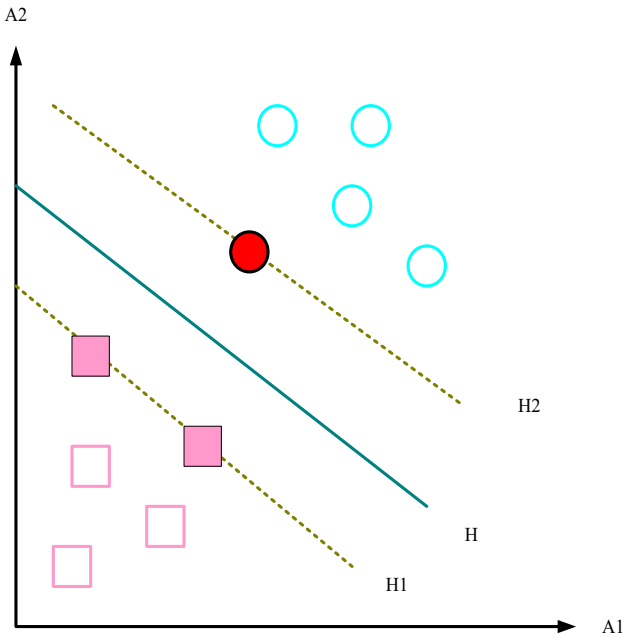


Fig. 3 SVM schematic

unique [18, 19]. When the hyperplane is determined, the distance from eigenvector a_i to the hyperplane can be approximated as $(w \cdot a_i + b)$. Whether the symbol of $(w \cdot a_i + b)$ is consistent with the symbol of the corresponding b_i determines whether the current classification is accurate. Therefore, the function interval from this point to the hyperplane is defined as γ_i, which is used to represent the accuracy and degree of accuracy of the current classification, which satisfies Formula 14:

$$\gamma_i = b_i(w \cdot a_i + b) \tag{14}$$

The function interval from the dataset to the hyperplane is the minimum value of the function interval from all sample points to the hyperplane, which is Formula 15:

$$\gamma_i = \min\{\gamma_1, \gamma_2, ..., \gamma_N\} \tag{15}$$

To make the interval and the plane have a one-to-one correspondence, the interval from the sample point to the plane is changed to Formula 16:

$$\gamma_i = b_i\left(\frac{w}{\|w\|} \cdot a_i + \frac{b}{\|w\|}\right) \tag{16}$$

In practical scenarios, there are often some special points in the dataset, and the sample set after ignoring these points can be regarded as linearly separable. Linear support vector machines are used to analyse such approximately linearly separable datasets.

Normalization is a method of simplifying calculations; that is, a dimensional expression is transformed into a dimensionless expression and becomes a scalar. Feature preprocessing refers to normalizing the extracted features.

In this paper, feature classification after fusion is realized by a multiclass support vector machine. Since each modality contains useful information for the current task, the features of different modalities may complement each other or cancel each other. Therefore, it is very important to choose an appropriate feature fusion method [20].

The eigenvectors of the same row in the feature matrix are composed of feature pairs $\alpha^T C_{\mathrm{ab}}\beta$, and each feature pair comes from two different modes. The purpose of the canonical correlation analysis is to find the mapping matrix as Formula 17:

$$\max(\alpha, \beta) = \frac{\alpha^T C_{ab}\beta}{\sqrt{\alpha^T C_{\mathrm{aa}}\alpha}\sqrt{\beta^T C_{\mathrm{bb}}\beta}} \tag{17}$$

where $C_{\mathrm{aa}} = A^T A$ represents the covariance matrix. The above optimization problem can be transformed into the solution of the matrix eigenvalue problem, that is, Formula 18:

$$\lambda^2\alpha_i = C_{\mathrm{aa}}^{-1}C_{\mathrm{ab}}C_{\mathrm{bb}}^{-1}\alpha_i \tag{18}$$

Canonical correlation analysis is based on linear space, and the nonlinear relationship between different modal characteristics cannot be obtained. Therefore, the kernel canonical correlation analysis (KCCA) method is proposed based on canonical correlation analysis, which adds nonlinear properties to the original canonical correlation analysis algorithm. The basic idea of KCCA is similar to the nonlinear support vector machine, which maps the original feature matrices a and b to high-dimensional space, namely kernel space A and B, and performs correlation analysis in the kernel space. The optimization function for kernel canonical correlation analysis is Formula 19:

$$\max(\alpha, \beta) = \frac{\alpha^T K_a K_b \beta}{\sqrt{\alpha^T K_a^2\alpha}\sqrt{\beta^T K_a^2\beta}} \tag{19}$$

The idea of kernel matrix fusion is to find a common subspace for two different modes, which can characterize the characteristics of the two modes to the greatest extent.

3.4 Design of sentiment tendency system based on deep learning

The problems caused by the manual collection of features in the traditional sentiment analysis system are as follows. The performance is not sufficient on small samples, and the effect is not good, but in this case, humans can recognize sentiments very well. The reason for this result is that the manual feature is not complete enough to fully express the sentiment features of the voice. The deep learning neural network has the ability of automatic feature learning, which can directly and automatically extract features

without much intervention or manual feature extraction. At the same time, the deep learning neural network also has its own classification function, so an end-to-end sentiment analysis system can be established.

This paper proposes an end-to-end model that is based on raw spectrogram input + deep neural network + SVM sentiment perception method. The deep learning model in this paper refers to the use of CNN-SVM, RNN-SVM, and the combination of CNN-SVM and RNN-SVM CRNN-SVM. The sentiment analysis system is shown in Fig. 4:

As shown in Fig. 4, the mobile terminal collects the voice signal on the mobile device and performs transcoding. When the user finishes speaking, the mobile terminal communication module sends the voice signal data to the server. The server scheduling module receives the request from the mobile terminal, allocates an ID identifier for this request, and applies it to the thread pool for an idle thread to start receiving voice data. After receiving the voice data, the scheduling module initiates a recognition request to the voice module. The voice module calls the idle thread to perform feature extraction on the received voice. After network identification, it returns, and the identification result is returned to the scheduling module. The scheduling module returns the result to the correct device according to the information in the returned recognition result and thus completes a speech sentiment tendency analysis request. The expression recognition module operates on the same principle. The scheduling module is a scheduling of all the previous modules as an entry to the pipeline.

Considering that people use mobile terminals as internet access devices most of the time in daily life, this paper designs a system based on mobile terminal devices. Due to the limited computing power of chips in mobile devices, a large number of deep network calculations are required to analyse speech sentiment. To reduce the system delay, this paper adopts the client/server structure, allowing the mobile device as the client to record and collect the user's sentimental voice signal. Then, it can be sent to the desktop computer as a server for neural network calculation, and

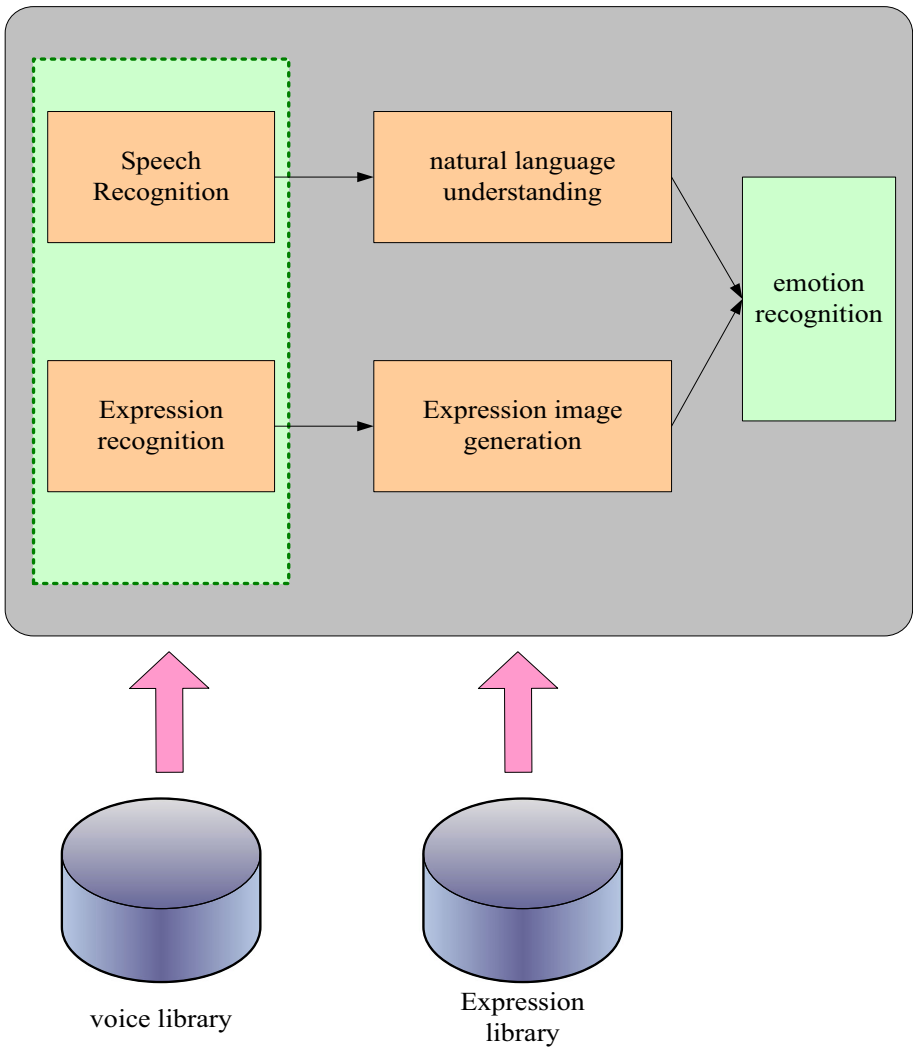


Fig. 4 Sentiment tendency analysis system

finally, the result is returned to the user terminal, as shown in Fig. 5.

As shown in Fig. 5, in this paper, the mobile phone is selected as the mobile device for the development and testing of the mobile terminal. Due to the limited time, the current application scenario of the system is to recognize the user's sentimental state during the conversation with the voice assistant and face recognition, and then the voice assistant responds appropriately according to the user's current sentimental state. The mobile terminal is developed using Java language, and the server terminal is developed using C + + language. C + + language is a statically typed, compile-time, cross-platform, irregular middle-level programming language that combines the characteristics of high-level and low-level languages.

The server is divided into two modules. The first module is the scheduling module, which is responsible for processing the voice sentiment analysis request from the mobile terminal. The application for an idle thread is used to receive the sentimental voice signal data sent by this request. In addition, the system supports large-scale concurrent task processing. The scheduling module processes requests from multiple mobile terminals and receives the recognition results returned by the module. The scheduling module needs to return the recognition result to the correct mobile device.

4 Sentiment tendency system experiment based on deep learning

4.1 Several model experiments based on deep learning

To verify the effectiveness of the various sentiment analysis methods proposed in this paper, experiments were carried out on three different multimodal sentiment databases, eNTRAFACE'05, RML, and AFEW6.0. In the experimental preparation stage, the video samples in the three databases are preprocessed, including facial expression recognition and speech extraction.

All the experimental parts of this article are based on the TensorFlow platform on the Ubuntu16.04.1TLS version. Since deep learning requires many computing resources, GPUs are often used to accelerate training. Its memory size is 22 GB, with a total of 4 GPU cores. This paper conducted experiments on the CNN-SVM, RNN-SVM, and CRNN-SVM models, as shown in Fig. 6:

As shown in Fig. 6, the smaller the step size is, the shorter the training time and the higher the accuracy on the test set. A stride of 75 times is not enough to show the best performance of CRNN-SVM. The reason for the above results is that when the step size is smaller, the number of samples in the training set generated is larger, resulting in a

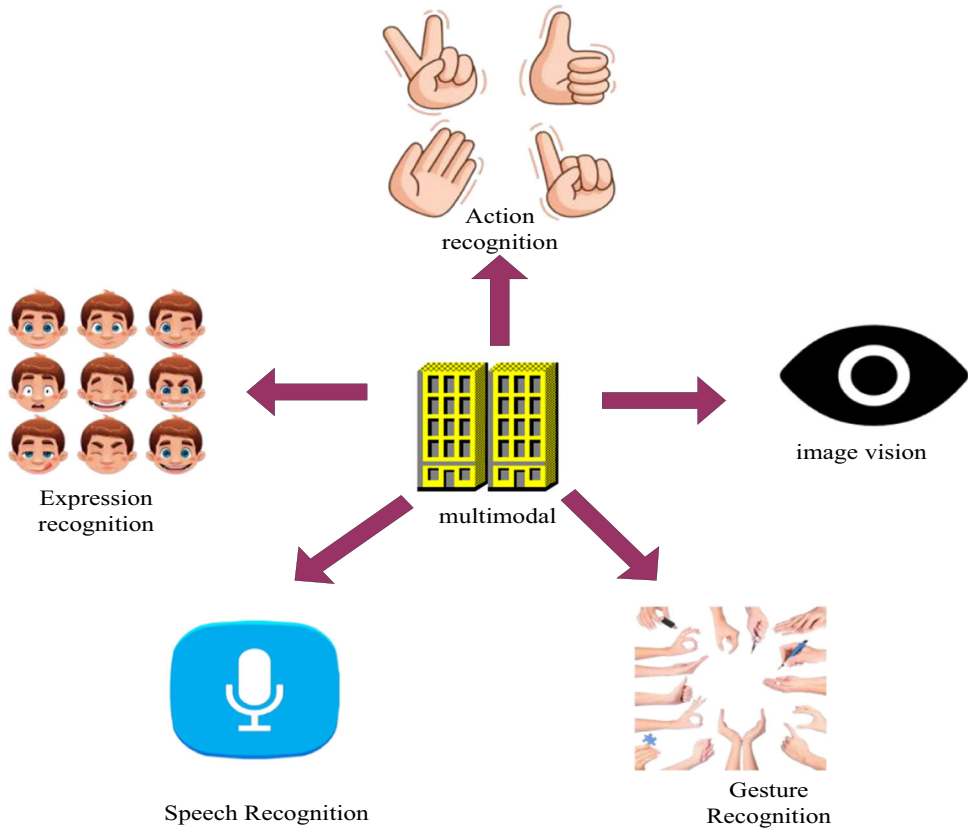


Fig. 5 Mobile terminal structure

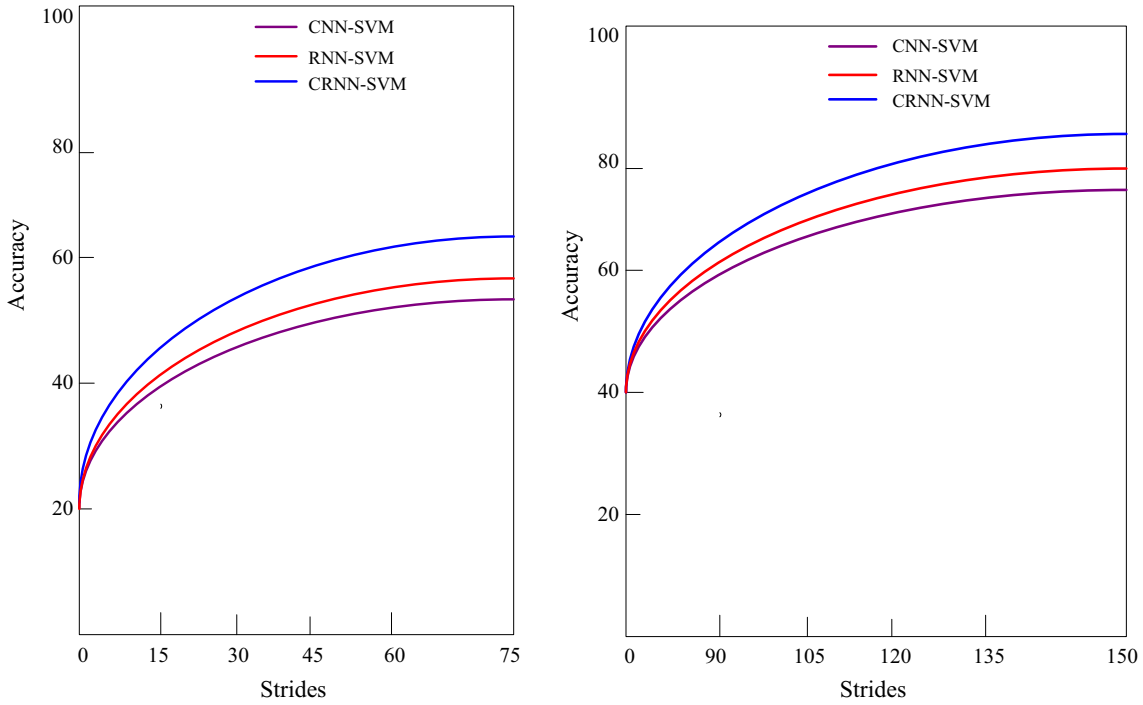


Fig. 6 Accuracy of CNN-SVM, RNN-SVM, and CRNN-SVM models with different steps

longer training time. Moreover, it leads to the lower possibility of overfitting the CRNN-SVM network, so when the step size is 150, the accuracy is the highest and the fitting is the best.

Using the test sample set to calculate the input features of the softmax classifier and inputting the calculated new features into the SVM, the classification results of CNN-SVM can be calculated. The structures of CNN-SVM and RNN-SVM are different except for the backend classifier, and the previous network structure is exactly the same. In this way, while iteratively calculating the effect of CRNN-SVM, this paper also obtains the accuracy comparison results of CNN-SVM and RNN-SVM, as shown in Table 1:

As shown in Table 1, CRNN-SVM has the highest sentiment tendency recognition rate in deep learning, so it is suitable to apply CRNN-SVM to the design of the sentiment tendency analysis system in this paper. The longitudinal comparison of CNN-SVM, RNN-SVM, and CRNN-SVM shows that the recognition effect of CNN-SVM and RNN-SVM is always smaller than that of CRNN-SVM. It shows that the features processed by CRNN are close to the degree of linear separability, so using CRNN-SVM can produce better results. The above results demonstrate the high efficiency of the CRNN-SVM sentiment perception model proposed in this paper.

Table 1 Accuracy comparison results of CNN-SVM and RNN-SVM

Strides	CNN-SVM	RNN-SVM	CRNN-SVM
8	66.5%	68.6%	89.8%
16	69.6%	69.0%	90.6%
24	72.4%	69.1%	91.9%
32	75.7%	70.0%	92.7%
36	76.8%	71.2%	93.5%

4.2 Multimodal sentiment tendency

All feature fusion methods in the experiment use a Gaussian kernel as the kernel function. In the method, according to the actual situation of the database, different weight ratios are selected. The specific parameter settings are shown in Table 2:

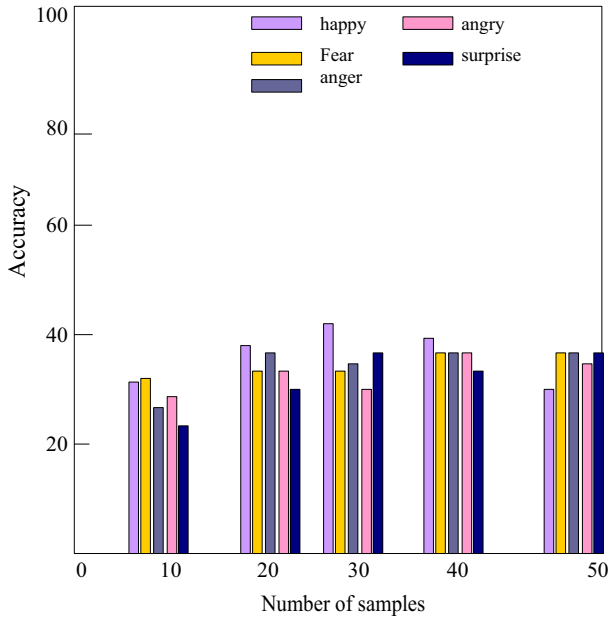
As shown in Table 2, the multimodal sentiment analysis in this paper includes facial expression recognition and speech sentiment analysis. In the multimodal sentiment database, the facial expression recognition method and the speech sentiment tendency analysis method with good

Table 2 Multimodal sentiment tendency analysis parameter settings

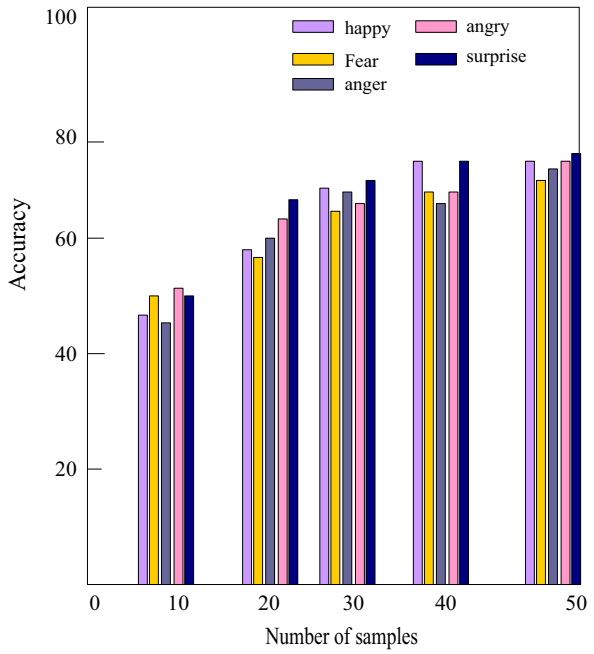
database	Facial expression weight	Voice weight
eNTRAFACE'05	0.7	0.19
RML	0.5	0.3
AFEW6.0	0.65	0.15

recognition effects are selected to verify the effectiveness of various fusion methods.

Another method for improving the generalization performance of the model is to increase the quantity of training data. For this reason, this paper selected 1 person who was best recognized by the system to supplement the corpus and expressions among the 15 recording personnel. The recognition results are shown in Fig. 7:



(a). Sentiment analysis effect of single-modal fusion



(b). Sentiment analysis effect of multi-modal fusion

Fig. 7 Sentimental tendency analysis effect of single-modal fusion and multimodal fusion

As shown in Fig. 7, the experimental results show that the sentiment tendency analysis effect of multimodal fusion is better than that of unimodal recognition. The overall recognition rate can be increased by 10% $\sim$ 20%, which shows that the study of multimodal fusion has practical significance. The experimental analysis compared various sentiment analysis methods, and the CRNN-SVM method has the best effect and universality. There are good results on the three databases used in the experiment. Enlarging the dataset and adding methods with different scene noise can improve the generalization ability of the model. Although it is not comparable to the tens of thousands of hours of training data for speech content recognition, it is believed that in the near future. With the increase in sentiment data, the effect of sentiment tendency analysis will have a breakthrough improvement.

4.3 System testing

In the development process, to ensure the performance and robustness of the system, it is necessary to carry out functional testing and performance testing of the system. The functional test ensures the normal operation of the functions of the major modules, which is completed by manual testing. The manual test ensures the normal function of the system by means of human operation without further elaboration. The following introduces the performance test of the system.

The next step is to test the system. In this paper, 50 texts in the CASIA database were simulated with 5 sentiment states, and a total of 500 pieces of data were obtained, which were then tested by the mobile phone voice sentiment analysis system. The model used on the mobile terminal was CRNN-SVM, and the accuracy of sentiment analysis of multimodal input speech is shown in Fig. 8:

As shown in Fig. 8, in the experiment, independent speech sentiment states are used. In this case, when the training sample size is small, it is not enough to contain all the speech quality characteristics. At this time, it is almost impossible to achieve independent speech sentiment recognition. The deep learning model requires a large quantity of data to support training the model, and the deep learning model has the ability to identify accurately.

This paper selects three speakers who are not in the training set and records a total of 600 sentences of sentiment test corpus in a conference room, office, and shopping mall environment. The test program simulates sending from the mobile phone to the server for identification and obtains the identification result. Tables 3 and 4 show the test results of the single-modal and multimodal recognition rates of the three speakers' sentiment orientation analysis:

As shown in Tables 3 and 4, the recognition rate of the multimodality method in the conference room scene is

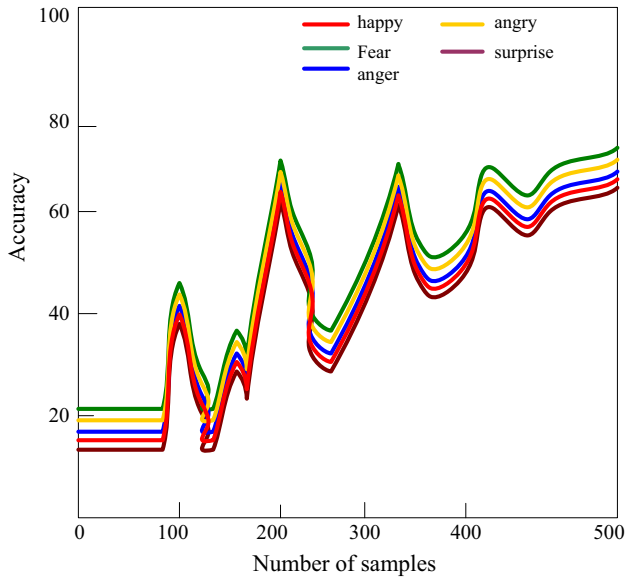


Fig. 8 The accuracy of sentiment analysis for multimodal speech

hardly improved. In the office and shopping mall environments, the improvement effect is significant, and the highest values are 90% and 95%, respectively, indicating that the multimodal method can improve the generalization ability of the model. Especially in a noisy environment, when the training data under certain conditions are small, the generalization ability of the single-modal model is relatively low, and the tolerance to environmental noise is small. The use of multimodal models can minimize errors in fewer data and avoid training a model being biased to a certain extreme, which would reduce the generalization ability of the network.

This paper simulates multiple mobile devices sending a large number of sentimental tendency analysis requests to the server at the same time, using 2,000 sentimental speech samples. The time taken from sending to receiving the recognition result is counted as shown in Table 5:

As shown in Table 5, the highest network transmission time is 62.5 ms, and the lowest is 57.6 ms; the highest processing time of the server is 125.1 ms, and the lowest is 112.6 ms. The overall average recognition time is up to 245.0 ms, which meets the system requirement that the processing time is less than 500 ms. During the experiment, the interval between consecutive requests received by the server was continuously reduced. When the interval is reduced to 80 ms, the system can work at full capacity without blocking. At this time, the size of the system thread pool is appropriate, that is, the maximum number of tasks processed meets the requirements. If the memory and computing power are sufficient, the server can fully meet the large-scale task processing requirements.

This paper then tests the long-term working stability of the system, using the test program to send a voice sentiment analysis request to the server every 80 ms, and this cycle is tested for 24 h. In these 24 h, a total of 1000 requests were sent, and the result of the server operation is shown in Fig. 9:

As shown in Fig. 9, the total number of "system busy" errors and "timeout processing" errors that occurred in the system accounted for a small proportion of the total requests, and within the expected range, the normal operation of the system accounted for the largest proportion. During this 24-h stress test, the server was running normally without any crashes. The system designed in this

Table 3 Single-modal sentiment analysis recognition rate

Number of samples	Meeting room %	Office %	Shopping mall environment %
50	67	70	75
100	69	73	73
150	72	75	78
200	70	74	77
250	71	72	76
300	68	71	79

Table 4 Multimodal sentiment analysis recognition rate

Number of samples	Meeting room %	Office %	Shopping mall environment %
50	73	89	92
100	70	87	95
150	71	90	93
200	74	86	89
250	75	88	91
300	66	90	94

Table 5 Time taken to identify results

Sample	Network transfer time	Server processing time	Overall recognition time
400	57.6 ms	112.6 ms	227.5 ms
800	58.4 ms	115.8 ms	228.9 ms
1200	59.7 ms	119.0 ms	230.8 ms
1600	60.3 ms	122.2 ms	237.6 ms
2000	62.5 ms	125.1 ms	245.0 ms

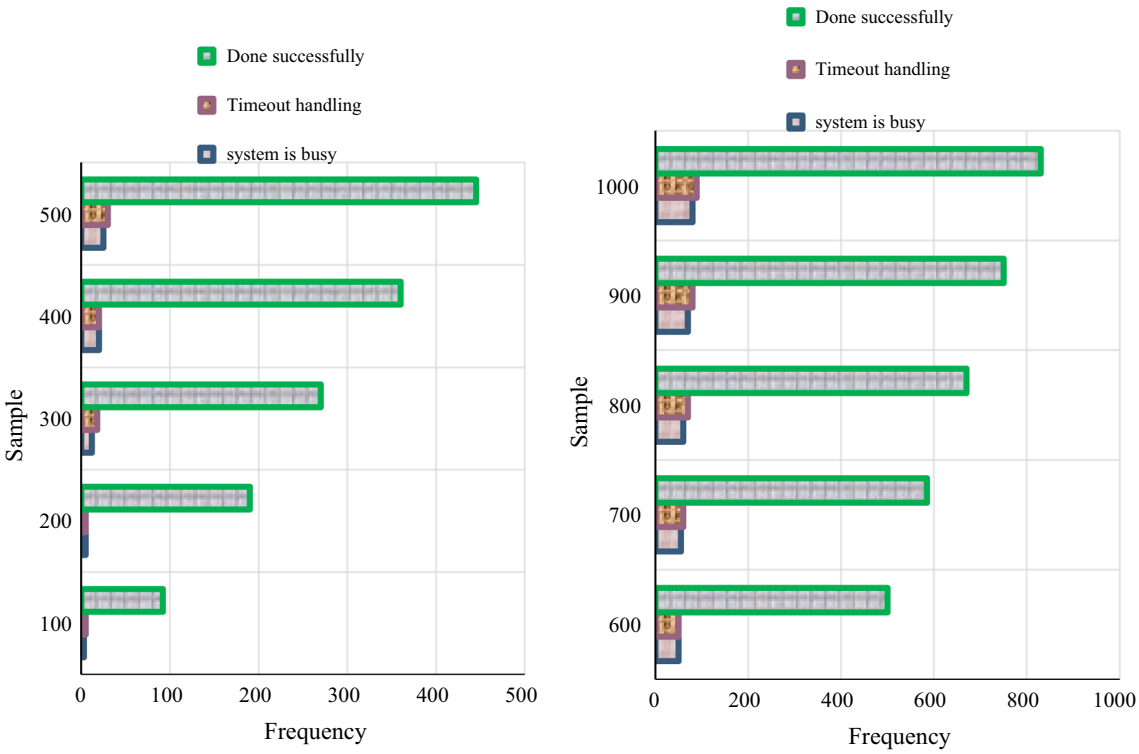


Fig. 9 The stability of the system for long-term operation

paper is a reliable and capable system for handling large-scale concurrent tasks.

5 Conclusion

Sentimental tendency refers to the sentiment polarity corresponding to sentiment words, focusing on positive, negative, and neutral sentiment. Sentimental tendencies generally refer to the psychological tendencies of people's views, evaluations, and opinions on events, objects, and social phenomena. The study of affective orientation has become a popular research direction in psychology and computer science. Generally, sentiment tendencies can be divided into positive, negative, and neutral. Sentiment orientation analysis is a branch of computer linguistics that involves knowledge of natural language processing, artificial intelligence, machine learning, and information retrieval. To better analyse people's sentiment tendencies, this paper proposes an AI deep learning method. The application of AI deep learning to sentiment tendency analysis is conducive to improving the recognition rate and analysis accuracy of sentiments. Based on AI deep learning, this paper simply designed a multimodal sentiment analysis system. Compared with the single-modal sentiment analysis system, the system can more accurately identify and analyse sentiments. In this paper, the system performance optimization method was tested, and remarkable results were achieved. Finally, to ensure system reliability, a long-time high-load stress test was carried out on the system. The results show that the system designed in this paper has high robustness and can be competent even for long-term high-load operation. In the experiment, due to the limitation of the sample, the result of the experiment may not be very scientific. To improve in future work, improvements should be made.

Funding This work was supported by the National Natural Science Foundation of China under Grant (No 0.61862061, 61563052, 61363064), 2018th Scientific Research Initiate Program of Doctors of Xinjiang University under Grant (No. 24470), Shaanxi Provincial Natural Science Foundation (No. 2020GY-093), and Shangluo City Science and Technology Program Fund Project (No. SK2019-83).

Data availability statement The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Young IF, Sullivan D, Stewart S, Palitsky R (2018) The existential approach to place: consequences for sentimental experience. J Environ Psychol 60:100–109
2. Yuan Z, Huang X, Zheng W, Zhen C, Zhao G (2018) Learning from hierarchical spatiotemporal descriptors for micro-expression recognition. IEEE Trans Multimed 20(11):3160–3172
3. Bertsch K, Roelofs K, Roch PJ, Ma B, Hensel S, Herpertz SC (2018) Neural correlates of sentimental action control in anger-prone women with borderline personality disorder. J Psychiatry Neurosci 43(3):161–170
4. Yan J, Zheng W, Cui Z, Tang C, Zhang T, Zong Y (2018) Multi-cue fusion for sentiment recognition in the wild. Neurocomputing 309:27–35
5. Alghifari MF, Gunawan TS, Kartiwi M (2018) Speech sentiment recognition using deep feedforward neural network. Indones J Electr Eng Comput Sci 10(2):554–561
6. Mittal T, Bhattacharya U, Chandra R (2020) M3er: Multiplicative multimodal sentiment recognition using facial, textual, and speech cues. In: Proceedings of the AAAI conference on artificial intelligence. 34(02): 1359–1367
7. Huan RH, Shu J, Bao SL (2021) Video multimodal sentiment recognition based on Bi-GRU and attention fusion. Multimed Tools Appl 80(6):8213–8240
8. Zhu XX, Tuia D, Mou L (2018) Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci Remote Sens Mag 5(4):8–36
9. Sun X, Wu P, Hoi S (2018) Face detection using deep learning: an improved faster RCNN approach. Neurocomputing 299:42–50
10. Elbamby MS, Perfecto C, Bennis M (2018) Towards low-latency and ultra-reliable virtual reality. IEEE Netw 32(2):78–84
11. Sunderhauf N, Brock O, Scheirer W, Hadsell R, Fox D, Leitner J (2018) The limits and potentials of deep learning for robotics. Int J Robot Res 37:405–420
12. He H, Wen CK, Shi J (2018) Deep learning-based channel estimation for beamspace mmWave massive MIMO systems. IEEE Wirel Commun Lett 7(5):852–855
13. Han J, Zhang D, Cheng G (2018) Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Process Mag 35(1):84–100
14. Palomino R, Low KB, Ji C (2021) Micro computed tomography analysis of four-way conversion catalysts using artificial intelligence-enabled image processing. Microsc Microanal 27(S1):1028–1029
15. Smith MA, Westerling-Bui T, Wilcox A (2021) Screening for bone marrow cellularity changes in cynomolgus macaques in toxicology safety studies using artificial intelligence models. Toxicol Pathol 49(4):905–911
16. Lee J (2020) Integration of digital twin and deep learning in cyber-physical systems: towards smart manufacturing. IET Collab Intell Manuf 38(8):901–910
17. Zheng Xu, Kamruzzaman MM, Shi J (2022) Method of generating face image based on text description of generating adversarial network. J Electron Imaging 31(5):051411
18. Balaganesh N, Muneeswaran K (2022) A novel aspect-based sentiment classifier using whale optimized adaptive neural network. Neural Comput Appl 34:4003–4012
19. Mittelstaedt JM, Wacker J, Stelling D (2019) Sentimental and cognitive modulation of cybersickness: the role of pain catastrophizing and body awareness. Hum Factors 61(2):322–336
20. Kaldewaij R, Koch SB, Zhang W, Hashemi MM, Klumpers F, Roelofs K (2019) Frontal control over automatic sentimental action tendencies predicts acute stress responsivity. Biol Psychiatry: Cognit Neurosci Neuroimaging 4(11):975–983