frontiers in Psychology

ORIGINAL RESEARCH
published: 22 October 2021
doi: 10.3389/fpsyg.2021.741665

Emotion Recognition of Chinese Paintings at the Thirteenth National Exhibition of Fines Arts in China Based on Advanced Affective Computing

Jing Li[1], Dongliang Chen[2], Ning Yu[1], Ziping Zhao[1] and Zhihan Lv[3]*

[1] College of Art, Qingdao Agricultural University, Qingdao, China, [2] College of Computer Science and Technology, Qingdao University, Qingdao, China, [3] Faculty of Arts, Uppsala University, Uppsala, Sweden

Today, with the rapid development of economic level, people's esthetic requirements are also rising, they have a deeper emotional understanding of art, and the voice of their traditional art and culture is becoming higher. The study expects to explore the performance of advanced affective computing in the recognition and analysis of emotional features of Chinese paintings at the 13th National Exhibition of Fines Arts. Aiming at the problem of "semantic gap" in the emotion recognition task of images such as traditional Chinese painting, the study selects the AlexNet algorithm based on convolutional neural network (CNN), and further improves the AlexNet algorithm. Meanwhile, the study adds chi square test to solve the problems of data redundancy and noise in various modes such as Chinese painting. Moreover, the study designs a multimodal emotion recognition model of Chinese painting based on improved AlexNet neural network and chi square test. Finally, the performance of the model is verified by simulation with Chinese painting in the 13th National Exhibition of Fines Arts as the data source. The proposed algorithm is compared with Long Short-Term Memory (LSTM), CNN, Recurrent Neural Network (RNN), AlexNet, and Deep Neural Network (DNN) algorithms from the training set and test set, respectively, The emotion recognition accuracy of the proposed algorithm reaches 92.23 and 97.11% in the training set and test set, respectively, the training time is stable at about 54.97 s, and the test time is stable at about 23.74 s. In addition, the analysis of the acceleration efficiency of each algorithm shows that the improved AlexNet algorithm is suitable for processing a large amount of brain image data, and the acceleration ratio is also higher than other algorithms. And the efficiency in the test set scenario is slightly better than that in the training set scenario. On the premise of ensuring the error, the multimodal emotion recognition model of Chinese painting can achieve high accuracy and obvious acceleration effect. More importantly, the emotion recognition and analysis effect of traditional Chinese painting is the best, which can provide an experimental basis for the digital understanding and management of emotion of quintessence.

Keywords: advanced affective computing, Chinese paintings, deep learning, the Thirteenth National Exhibition of Fines Arts in China, emotion recognition

OPEN ACCESS

Edited by:
Yizhang Jiang,
Jiangnan University, China

Reviewed by:
Mary Nina Wang,
University of Malaya, Malaysia
Yuan Fu,
Tianjin University, China

****Correspondence:***
Ning Yu
yuningluck@126.com

Specialty section:
This article was submitted to Emotion Science, a section of the journal Frontiers in Psychology

Received: *15 July 2021*
Accepted: *24 September 2021*
Published: *22 October 2021*

Citation:
Li J, Chen D, Yu N, Zhao Z and Lv Z (2021) Emotion Recognition of Chinese Paintings at the Thirteenth National Exhibition of Fines Arts in China Based on Advanced Affective Computing. Front. Psychol. 12:741665. doi: 10.3389/fpsyg.2021.741665

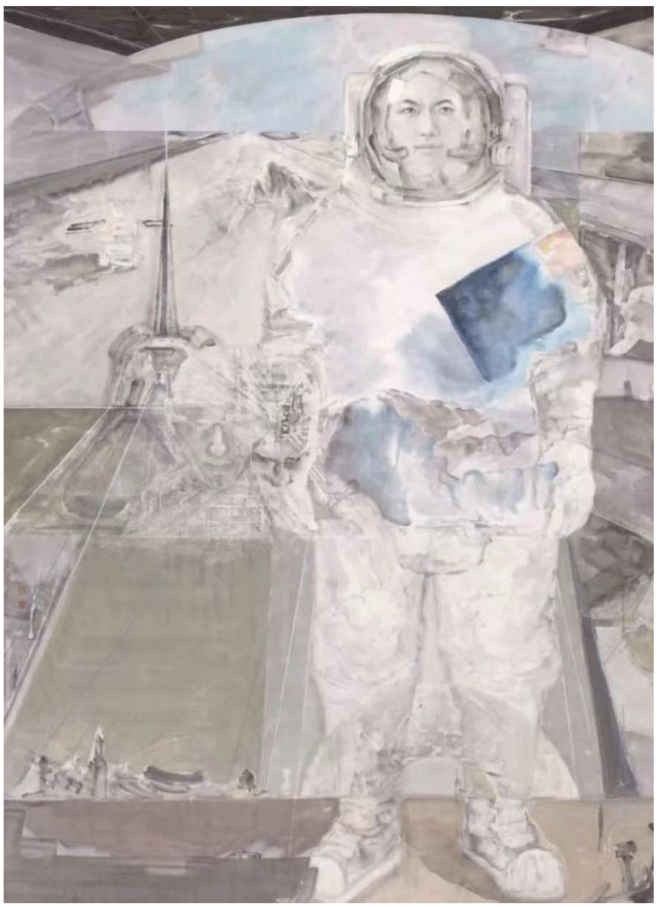
INTRODUCTION

Now that the world comes to the big data era, Artificial Intelligence (AI) has impacted people's lives profoundly; the traditional interpersonal interactions begin to shift to human-computer emotional interactions. In the meantime, people put forward higher requirements for esthetic appreciations than in the past. The performance of simple painting techniques has been unable to meet the emotional needs of artists, more and more artists focus on the emotional expression of works. In this case, how to understand the emotional expression of works of art is particularly important. More and more researchers are committed to the emotional expression of works of art.

The works of the Thirteenth National Exhibition of Fines Arts are close to the daily life of the Chinese people. It reflects the new changes in the development of modern society and the achievements of China's development in the new era, and shapes and depicts the spirit of the Chinese nation in the new era. The Chinese paintings in this exhibition have the following emotional descriptions: Li Encheng's *Fanghua* shows the posture of natural mountains and fields, praising the unknown and selfless workers in all walks of life. *Mission* embodies the patriotic spirit of heroism and model practitioners—fire fighters. *Pursuing the Dream of Space* shows the astronauts' exploration of the secrets of the universe, the endless Chinese spirit, and Chinese power. The selected works of Chinese painting in the Thirteenth National Exhibition of Fines Arts reveal that, the themes consciously highlight the spirit of the times, depict all kinds of people's livelihood, the Chinese mountains and rivers are thriving, and the painting themes show a new fashion of the times.

Fanghua by Li Encheng

Mission by Li Yuwang



Pursuing the Dream of Space by Sun Chunlong

It may be challenging to capture the emotions expressed in these paintings accurately. Therefore, using Advanced Affective Computing to analyze the emotions expressed in these paintings can offer an opportunity for the general public to appreciate art. However, the result may be contrary to the user's actual emotion applying the traditional single model for Chinese painting sentiment analysis. Regarding the complementary associations of various modality information, multimodal emotion recognition has gradually received widespread attention to thoroughly utilize the present modality data to capture information in and between the modalities (He and Zhang, 2018). Users pose new challenges to emotion recognition while they employ different modalities to portray their real-time emotions and make their current emotions vivid.

As for image data such as Chinese paintings, features such as SIFT (Scale-Invariant Feature Transform) operator vectors can be obtained. Original feature data of different modalities have their unique structural characteristics belonging to different feature spaces. Hence, noise modality may also be mixed into them during feature extraction (Zhang et al., 2020). Traditional emotion recognition algorithms may increase the calculation time and space cost, resulting in disturbed recognition outcomes. As of now, many large-scale, high-quality datasets are proposed, such as ImageNet, Deep Learning (DL) approaches have made breakthrough achievements in image recognition. Compared with artificially designed features, the unique multi-layer structure of DL enables it to learn the in-depth features that gradually transits from universal, low-level visual features (such as edges and textures) to high-level semantic representations (such as the torso and head); the higher the level, the stronger the expression (Wang et al., 2019; Chen et al., 2020). The hierarchical nature of the deep features provides a practical means to bridge the semantic gap and understand human affects in paintings (Liu et al., 2021) .

To sum up, under the trend of the rapid development of AI, using advanced affective technology to identify the emotion of Chinese painting with multiple feature spaces has become the internal driving force for people to understand artistic emotion, which has great practical significance for the inheritance of Chinese traditional culture. Therefore, the innovation is to solve the problem of "semantic gap" in the emotion recognition task of images such as traditional Chinese painting, select and improve the AlexNet algorithm on the basis of deep learning theory to further improve its performance. The study also adds chi square test to solve the data redundancy and noise problems in various modes such as Chinese painting. Additionally, a multimodal emotion recognition model of Chinese painting based on improved AlexNet neural network and chi square test is designed to provide theoretical support for the digital understanding and development of emotion in Chinese quintessence.

RELATED WORKS

Trend of Intelligent Emotion Recognition

Emotion capture is the foundation of daily communication. Multi-party cooperation in all professions and trades is inseparable from emotion analysis and emotion recognition. Jain et al. (2018) proposed a hybrid DNN (Deep Neural Network) to recognize emotions in face images. They tested this hybrid DNN on two public datasets and harvested satisfactory experimental results (Jain et al., 2018). Avots et al. (2019) captured and classified facial expressions using SVMs (Support Vector Machines). They utilized Viola-Jones facial recognition algorithm and CNN (AlexNet) emotion classification algorithm to find human faces in keyframes. Ultimately, the proposed algorithm was validated (Avots et al., 2019). Hwang et al. (2020) put forward a novel emotion recognition approach based on CNN (Convolutional Neural Network) while preventing local information loss. Lastly, visualization proved that this approach outperformed other Electroencephalography (EEG) feature representation models based on standard features, proving its effectiveness (Hwang et al., 2020). Wei et al. (2021) designed an algorithm to extract keyframes from videos based on emotion saliency estimation regarding the current status of video emotion recognition. Keyframes could be extracted to avoid the influence of the emotion-independent frames on the result by estimating the emotional saliency of the video frame. Through simulation, the designed algorithm could improve the performance of video emotion recognition, outperforming the present user-generated video emotion recognition (Wei et al., 2021).

Advanced Affective Computing for Multimodal Analysis

Hammad et al. (2018) put forward a secure multimodal biometric system based on CNN and a QG-MSVM (Q-Gaussian Multi-Support Vector Machine) based on different fusion levels. They applied this internal fusion system to combine the biological characteristics to generate a biological characteristic template. Results demonstrated that the proposed system could provide higher efficiency, robustness, and reliability than the present multimodal authentication system (Hammad et al., 2018). Ćosić et al. (2019) enhanced the process of ATC (Air Traffic Controller) selection based on the traditional ATC psychophysiological data measurement, including complicated physiological, eye volume, and voice measurements and appropriate metrics, as well as the ability to measure the multimodality in particular stimulus tasks. They comprehensively analyzed the multimodal features of this method under different experimental conditions and found it pretty advantageous (Ćosić et al., 2019). Yuan et al. (2020) designed and implemented the MSNVRFL (Multimodal Sensing Navigation Virtual and Real Fusion Laboratory) regarding the importance of virtual experiments in HCI (Human-Computer Interaction). MSNVRFL was equipped with a set of experimental devices with cognitive functions, in which a multimodal chemical experiment fusion algorithm was explored, validated, and applied (Yuan et al., 2020). Walia and Rohilla (2021) reviewed the various biological activity detection technologies based on multimodal biometric systems. After analysis and discussion, they proposed a new classification approach and finally proved its effectiveness (Walia and Rohilla, 2021).

Apparently, most works about image sentiment analysis are based on face images or sequences; in contrast, the sentiment analysis for images such as Chinese paintings is still a great challenge. Works about intelligent Chinese painting recognition have been reported; nevertheless, they focused on identifying the author of the painting and analyzing the work style, with very little research on the emotions shown by Chinese paintings. Intelligent algorithms, such as DL, have not played a significant role in emotion recognition. In this case, DL approaches are employed in the present work to identify and analyze the emotions expressed in Chinese paintings, which is of great significance for the in-depth understanding of the emotions of Chinese paintings subsequently.

MULTIMODAL EMOTION RECOGNITION AND ANALYSIS OF CHINESE PAINTINGS AT THE THIRTEENTH NATIONAL EXHIBITION OF FINES ARTS IN CHINA BASED ON DL APPROACH

The Thirteenth National Exhibition of Fines Arts vividly shows the spirit of the Chinese nation in the new era and the happy life of the people. A DL-based multimodal emotion recognition model for Chinese paintings is proposed based on image data of Chinese paintings at the Thirteenth National Exhibition of Fines Arts in China. Subsequently, its performance is validated through comparative simulation experiments.

Demand for Chinese Painting Emotion Recognition

Chinese paintings pay attention to image modeling. It cannot blindly imitate the objective reality, nor can it be purely fabricated, but pursues the artistic state of "like and not like." Based on respecting the objective images, painters integrate their feelings into the objective things. In this way, they can not only express their feelings in the works, but also make the images have emotion and vitality. Emotion is the driving force of Chinese painting creation (Bi et al., 2019; Yang et al., 2021). Traditional algorithms to recognize human affects in Chinese paintings often combine art theory and computer vision, practicing artificially designed features and statistical ML (Machine Learning) approaches to recognize the emotional responses evoked by Chinese paintings. The emotional response of Chinese paintings is presented in **Figure 1**. However, the large-scale, high-quality labeled datasets, such as ImageNet, make traditional ML approaches incapable of training deeper DL models, leading to severe overfitting.

Regularly, the large-scale dataset distribution used for pre-training is very different from the dataset distribution of the target task. For instance, the ImageNet dataset is dominated by natural images (such as animals and natural scenes). In contrast, image sentiment classification datasets such as Chinese paintings are completely composed of painting elements. Thus, if the pre-training weights of the ImageNet dataset are directly transferred, under-fitting caused by cross-domain transfer will occur. Meanwhile, a large-scale, image emotion recognition dataset has not been established yet. In this case, how to resolve under-fitting becomes a hot topic.

An emotion recognition model for images such as Chinese paintings is established based on DL approaches to bridge the gap between the universal, low-level visual features and high-level emotional semantics and address the small sample issue of emotion recognition datasets. Afterward, DL overfitting in the case of a small dataset gets improved based on TL (Transfer Learning). In the meantime, a two-layer TL strategy is proposed for emotion recognition of Chinese paintings to resolve under-fitting during cross-domain transfer.

DL Approaches for Chinese Painting Emotion Recognition

While appreciating Chinese paintings, people often evaluate art pieces regarding artistic conception, charm, and drawing techniques. If emotions in Chinese paintings can be decomposed and analyzed quantitatively, people will understand the way Chinese paintings express emotions and thus better appreciate the beauty of Chinese paintings. Besides, emotion visualization can intuitively demonstrate the digital laws of emotions behind Chinese paintings, greatly enlarging the effect and quality of cultural and artistic learning, analysis, and training (Wang et al., 2017). CNN, RNN (Recurrent Neural Network), and LSTM (Long Short-Term Memory) are standard image recognition algorithms. In the present work, CNN is selected for emotion recognition. CNN is a feedforward neural network, which often contains multiple layers in different types, such as convolutional layers, fully connected layers, and pooling layers (Haberl et al., 2018; Kim et al., 2020). **Figure 2** visualizes the procedures of CNN processing and recognizing Chinese paintings.

Normally, CNN performs convolution operations in multiple dimensions. Suppose the input is a 2D matrix I; in that case, there is a 2D kernel K satisfying the following equation:

$$S(i,j) = (I \cdot K)(i,j) = \sum_m \sum_n I(m,n)\, K(i-m, j-n) \quad (1)$$

In (1), *(i, j)* describes the matrix dimension, and *(m, n)* refers to the matrix order. Convolution can be exchanged and equivalently written as the following equation:

$$S(i,j) = (I \cdot K)(i,j) = \sum_m \sum_n I(i-m, j-n) K(m,n) \quad (2)$$

The flipped convolution kernel, as opposed to the input, gives interchangeability features to the convolution operation; the input index is increasing, while the kernel index is decreasing. The only goal of kernel flipping is to achieve exchangeability; although this feature is helpful in proofs, it is not that significant to neural networks. Instead, many neural network libraries possess a function called CC (Cross-Correlation) (Lin et al., 2018), almost the same as the convolution operation but cannot flip the kernel:

$$S(i,j) = (I \cdot K)(i,j) = \sum_m \sum_n I(i+m, j+n) K(m,n) \quad (3)$$

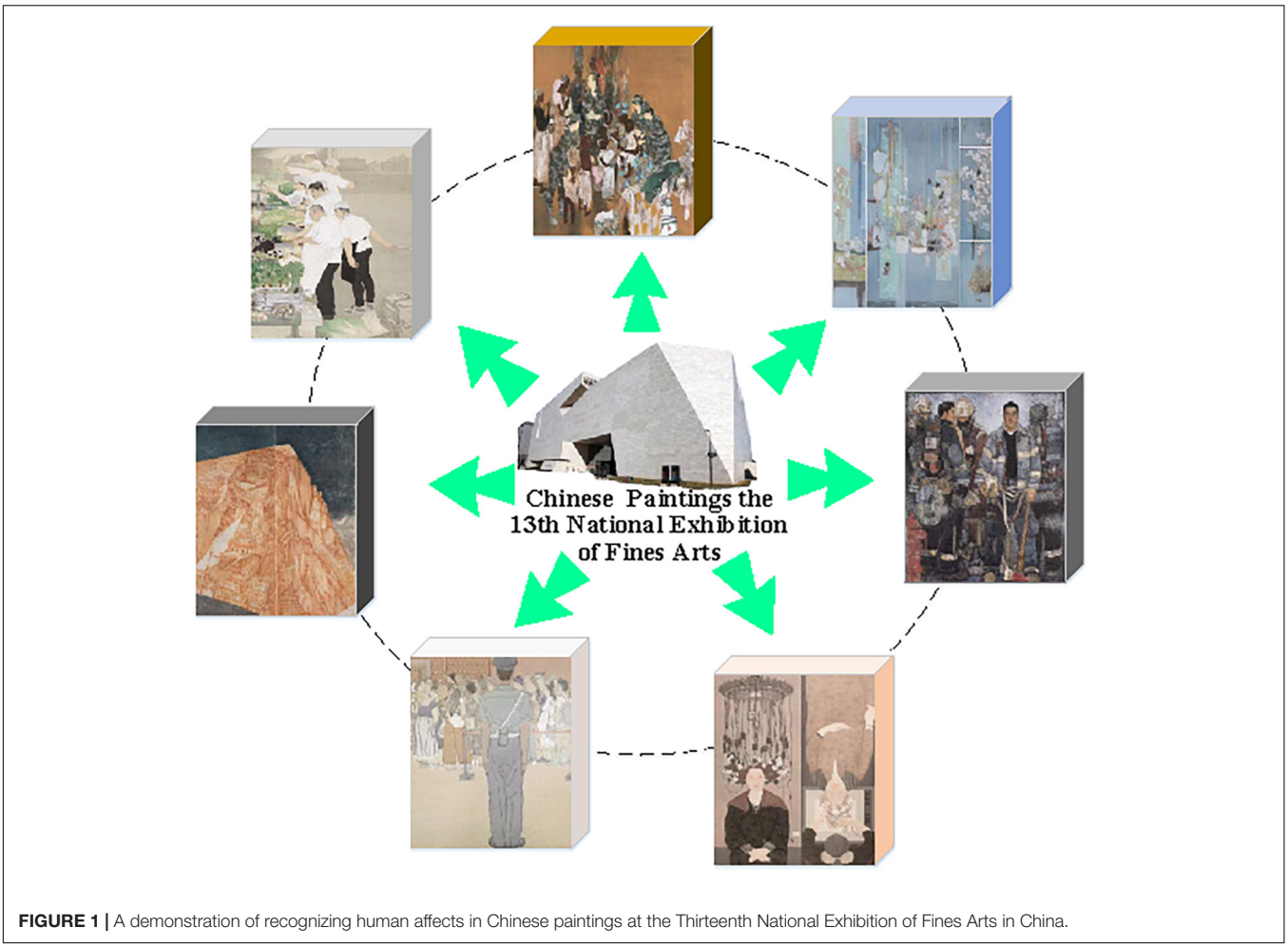


FIGURE 1 | A demonstration of recognizing human affects in Chinese paintings at the Thirteenth National Exhibition of Fines Arts in China.

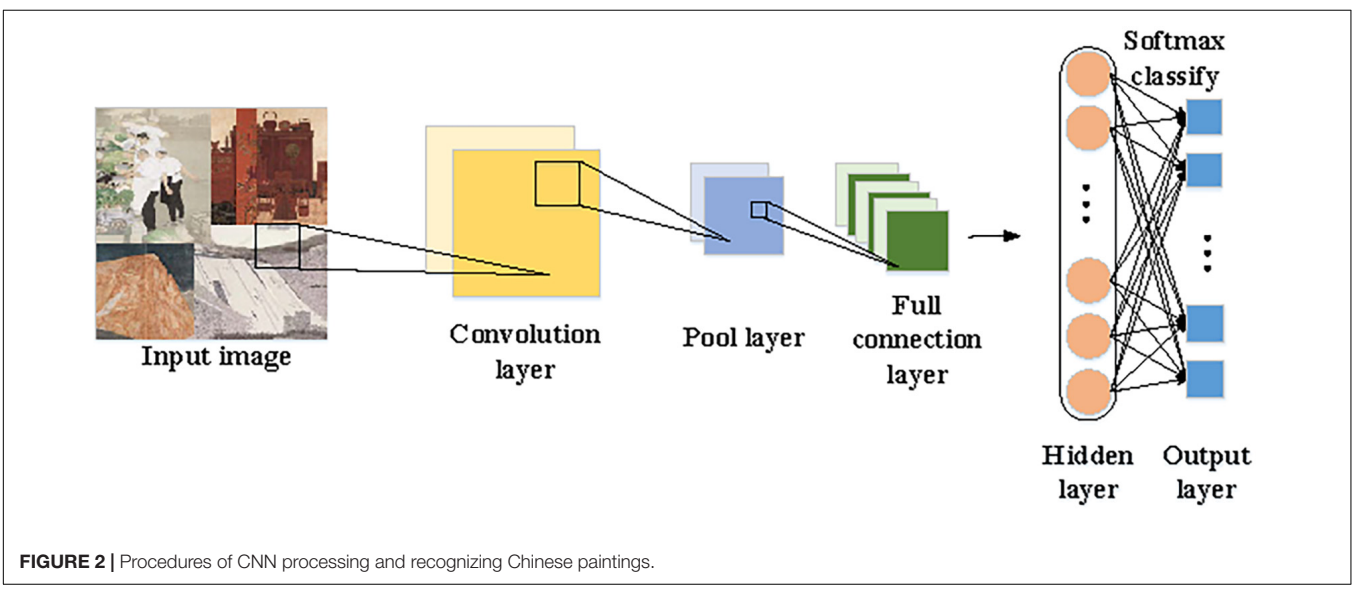


FIGURE 2 | Procedures of CNN processing and recognizing Chinese paintings.

Convolutional neural network classifies pixels of the original Chinese paintings before the up-sampling operation that continuously scales the image to the original size. The final output is the up-sampled image, so that each pixel in the output image can be forecasted. In particular, the maximal image pixel is found in all the images finally obtained. AlexNet has more layers and

stronger learning capability among various CNNs; thus, it is selected in the present work to reduce the calculation amount and strengthen the generalization performance of the algorithm (Jia et al., 2019). Suppose that $A_i^{(l)}$ denotes the output of the l-th convolutional layer in CNN, $A^{(0)}$ describes the Chinese painting input; in that case, its $i\left(1 \le i \le M^{(l)}\right)$-th feature map can be calculated through:

$$A_i^{(l)} = \sigma\left(\sum_{j=1}^{M^{(l-1)}} A_j^{(l-1)} \otimes W_{i,j}^{(l)} + b_i^{(l)}\right) \tag{4}$$

In (4), $M^{(l)}$ refers to the total number of feature maps of the l-th convolutional layer, W represents the weight of the convolution kernel, b is the bias, $\otimes$ signifies the convolution operation, and $\sigma(\cdot)$ demonstrates ReLU (Rectified Linear Unit) (Gu et al., 2019). ReLU can fix vanishing and exploding gradient and enhance the network's expression. In the pooling layer, standard pooling strategies are Max-Pooling and Average-Pooling. AlexNet often employs the Max-Pooling strategy. After the convolutional layer and the pooling layer map the original data to the hidden layer characteristic space, multiple fully connected layers are connected to map the learned feature representation to the sample's label space. AlexNet has three fully connected layers; their equations are:

$$A_i^{(6)} = \sigma\left(\sum_{j=1}^{n^{(5)}} W_{i,j}^{(6)} \times F_j\left(A^{(5)}\right) + b_i^{(6)}\right) \tag{5}$$

$$A_i^{(7)} = \sigma\left(\sum_{j=1}^{n^{(6)}} W_{i,j}^{(7)} \times A^{(6)} + b_i^{(7)}\right) \tag{6}$$

$$A_i^{(8)} = \varphi\left(\sum_{j=1}^{n^{(7)}} W_{i,j}^{(8)} \times A^{(7)} + b_i^{(8)}\right) \tag{7}$$

In (5) $\sim$ (7), $n^{(l)}$ refers to the number of neurons in the l-th layer, $F(\cdot)$ refers to the tiling operation in which the result of the last convolutional layer is expanded into a 1D vector, $\varphi(\cdot)$ signifies SoftMax that predicts the probability that the input vector belongs to each emotion category, and its equation is:

$$\varphi_j(X) = \frac{e^{X_j}}{\sum_k e^{X_k}} \tag{8}$$

In (8), j represents the j-th element of the input vector X, and k denotes the total number of sample categories. Furthermore, AlexNet's convolutional layer is improved. The operation of "local normalization before pooling" is advanced to "pooling before local normalization." This improvement brings two benefits. First, the generalization ability of AlexNet can be enhanced while the over-fitting can be weakened, which greatly shortens the training time. Second, pooling before local normalization can retain useful information, weaken redundant information, and accelerate the convergence of the proposed algorithm, highlighting the superiority of overlapping Max-Pooling. **Figure 3** presents the procedures of the improved AlexNet processing and recognizing Chinese paintings.

Removing redundant information from feature data is a necessary step before an emotion recognition task, especially when the amount of original modality data is particularly huge. Traditionally, text information and image information can be extracted from images for emotion recognition. However, emotion recognition of Chinese paintings is a binary classification task. Only using image information can accurately determine the negativity and positivity of the user's current emotion. Hence, image information occupies more weight than text information. If text information appears in this case, it will be regarded as subordinate redundant information. A process to remove redundant information in the modality feature, retain the weighted feature in the emotion recognition task, and reduce the calculation cost and resource wastes is extremely meaningful and remarkable. There are three well-known feature selection methods: packaging, embedding, and filtering. The packaging method comprises a learning algorithm that can evaluate the accuracy performance of a subset of candidate features, providing better solutions than the other two methods. SVM-RFE (Support Vector Machine-Recursive Feature Elimination) is a standard packaging method (Spoorthi et al., 2018; Nalepa et al., 2019; Sukumarana and Mb, 2019). Suppose a given training sample set as $\{x_i, y_i\}, x_i \in R^d, y_i \in \{-1, 1\}, i = 1, \ldots, n$; in that case, the decision function of linear SVM is:

$$f(x) = w * x + b \tag{9}$$

In (9), w represents the weight, and b denotes the bias term. The boundary M is proved to be only $2/|w|$. Hence, the maximum boundary is equivalent to minimizing $||w||^2$ under constraints. The dual form of the LaGrange problem can be expressed as:

$$L_D = \sum_{i=1}^{n} \partial_i - \frac{1}{2}\sum_{j=1}^{n} \partial_i \partial_j y_i y_j x_i x_j \tag{10}$$

In (10), ∂_i refers to the LaGrange multiplier. The solution to ∂_i can be calculated by maximizing L_D under constraints $\partial_i \ge 0$ and $\sum_{i=1}^{n} \partial_i y_i = 0$. Samples corresponding to non-zero ∂ are called support vectors. The weight vector w can be obtained by the following equation:

$$w = \sum_{i=1}^{n} \partial_i y_i x_i \tag{11}$$

Then, the ranking criterion of the pixel feature k of Chinese paintings is the square of the k-th element of w:

$$J(k) = w_k^2 \tag{12}$$

SVM-RFE trains all SVM features to calculate their weights, which is pretty cumbersome. Chi Square Test and information gain are two widespread filtering methods. During feature extraction, Chi Square Test measures the dependency between features and categories. The higher the score, the more dependent the categories are on the given features. Thus, features with lower scores have less information and should be deleted. Suppose

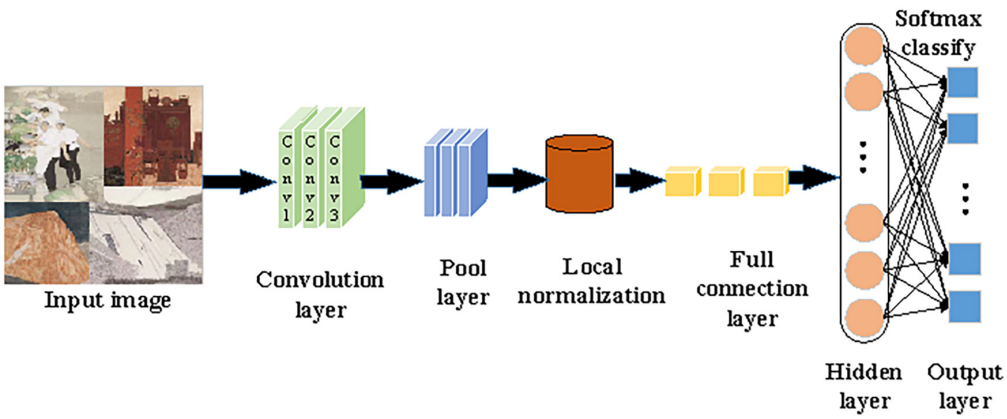


FIGURE 3 | Procedures of the improved AlexNet processing and recognizing Chinese paintings.

that a feature in a Chinese painting is independent of the final classification; in that case, the Chi Square Test can be defined as:

$$CHI(t, c_i) = \frac{N \times (AD - BC)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (13)$$

$$CHI_{\max}(t) = \max\left(CHI(t, c_i)\right) \quad (14)$$

In (13) and (14), A refers to the number of documents with feature t and belonging to category c_i , B describes the number of documents with feature t but not belonging to category c_i , C refers to the number of occurrences of c_i without t, D signifies the frequency where neither c_i nor t appears, and N refers to the total number of instances in the document set, $N = A+B+C+D$. If t and c_i are independent, the *Chi* statistic will be zero.

Information gain measures the information obtained after the feature values in the document are known. The higher the information gain, the better the ability to distinguish different categories. The information can be calculated by capturing the uncertainty entropy of the probability distribution of a given category. Suppose there are m categories: $C = \{c_1, c_2, \cdots, c_m\}$; in that case, the following equation can be deduced based on entropy:

$$H(C) = -\sum_{i=1}^{m} p\left(c_i\right) \log_2 p\left(c_i\right) \quad (15)$$

In (15), $p\left(c_i\right)$ denotes the probability of how many documents are in category c_i. Suppose that the attribute A in Chinese paintings has n different values: $A = \{a_1, a_2, \cdots, a_m\}$; in that case, the entropy after examining attribute A can be defined as follows:

$$H(C|A) = \sum_{j=1}^{m} \left(-P\left(a_j\right) \sum_{i=1}^{m} p\left(c_i|a_j\right) \log_2 p\left(c_i|a_j\right) \right) \quad (16)$$

In (16), $P\left(a_j\right)$ suggests the probability of how many documents contain the attribute value a_j , and $p\left(c_i|a_j\right)$ describes the probability of the number of documents containing the attribute value a_j in category c_i . Based on the above definition, the information gain of an attribute is exactly the difference between the entropy values before and after examining the attribute, which can be expressed as:

$$IG(A) = H(C) - H(C|A) \quad (17)$$

Therefore, during sentiment analysis, the filtering method can select a subset of features from the original features, use statistical measures to rank the available features, and automatically filter out those features whose scores are lower than a predetermined threshold. More importantly, it can assess feature subsets without learning algorithms. In short, this method is easy to design and does not require substantial computing resources, showing significant advantages for large datasets.

The Multimodal Emotion Recognition Algorithm for Chinese Paintings at the Thirteenth National Exhibition of Fines Arts in China Based on Chi Square Test and Improved AlexNet

A multimodal emotion recognition algorithm for Chinese paintings is designed based on improved AlexNet and Chi Square Test. In particular, because it has more layers and stronger learning capability than other CNNs, AlexNet is chosen in the present work and gets improved to solve the "semantic gaps" in recognizing emotions from images such as Chinese paintings. Furthermore, Chi Square Test removes data redundancy and data noise in each modality of Chinese paintings and captures the internal connections among the modalities. This algorithm not only addresses the above concerns but also saves calculation costs and improves emotion recognition accuracy. **Figure 4** below explains its procedures.

The proposes algorithm obeys the following two principles: (1) thoroughly utilizing different modality data, avoiding any wastes, and (2) choosing the most relevant features in high-dimensional spaces to discover the internal connections among features. It can provide more excellent classification accuracy than other algorithms.

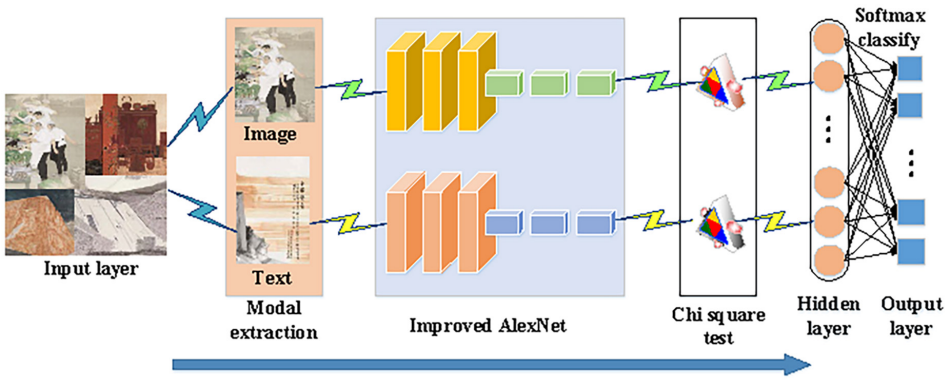


FIGURE 4 | Procedures of the proposed multimodal emotion recognition algorithm based on Chi Square Test and improved AlexNet processing Chinese paintings at the Thirteenth National Exhibition of Fines Arts in China.

During training, one local normalization layer is attached after AlexNet's pooling layer to standardize the feature map $c_t^l(i,j)$:

$$c_t^l(i,j) = a_t^l(i,j) / \left(k + \alpha \sum_{\max(0,t-m/2)}^{\min(N-1,t+m/2)} \left(a_t^l(i,j) \right)^2 \right)^{\beta} \quad (18)$$

In (18), k, α, β, m are hyperparameters valuing 2, 0.78, 10^{-4}, and 7, respectively, and N is the total number of convolution kernels in the l-th convolutional layer. To prevent "gradient dispersion" (Xu and Li, 2021) ReLU is employed to activate the convolution output $S_t^l(i,j)$:

$$y_t^l(i,j) = f\left(S_t^l(i,j)\right) = \max\left\{0, S_t^l(i,j)\right\} \quad (19)$$

In (19), $f()$ represents ReLU. To avoid over-fitting in the fully connected layer, the dropout parameter is set to 0.5.

While improving the generalization ability, neurons C^{ll} in the fully connected layer are discarded and output, $r_j^l \sim bernoulli(dp)$, $\tilde{C}^l = r^l C^l$; in that case, the i-th neuron input in the fully connected layer Z_i^{l+1} is $W_i^{l+1}\tilde{C}^l + b_i^{l+1}$, where the i-th neuron input in the next fully connected layer C_i^l is $f\left(Z_i^l\right)$, namely $\max\left\{0, Z_i^l\right\}$. Eventually, the input q^i of the i-th neuron in the fully connected layer can be obtained:

$$q^i = soft\max(Z_i^8) = \frac{e^{Z_i^8}}{\sum_{j=1}^{12} e^{Z_i^8}} \quad (20)$$

Here, the cross-entropy loss function suitable for classification is taken as the algorithm's error function, and the equation is:

$$Loss = \sum_{i=1}^{K} y_i \cdot \log\left(p_i\right) \quad (21)$$

$$p_i = \frac{\exp\left(\tilde{y}_i\right)}{\sum_{i=1}^{K} \exp\left(\tilde{y}_j\right)} \quad (22)$$

In (21) and (22), N signifies the number of categories, y_i represents the actual category distribution of samples, $\tilde{y}_i$ describes the network output, and p_i denotes the result after the SoftMax classifier. SoftMax's input is an N-dimensional real number vector, denoted as x. Its equation is:

$$\xi(x)_i = \frac{e^{x_i}}{\sum_{n=1}^{N} e^{x_i}}, i = 1, 2, ..., N \quad (23)$$

Essentially, SoftMax can map an M-dimensional arbitrary real number vector to an M-dimensional vector whose values all fall in the range of (0,1), thereby normalizing the vector. To reduce the computational amount, the output data volume is reduced to 2^8 through μ companding conversion, that is, $\mu = 255$, thereby improving the algorithm's forecasting efficiency.

$$f(x_t) = sign(x_t) \frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}, |x_t| < 1 \quad (24)$$

The proposed algorithm is trained through learning rate updating using the polynomial decay approach (Poly) (Cai et al., 2020). The equation is:

$$init\_lr \times \left(1 - \frac{epoch}{\max\_epoch}\right)^{power} \quad (25)$$

In (26), the initial learning rate $init\_lr$ is 0.0005 (or $5e^{-4}$), and power is set to 0.9. The Weighted Cross-Entropy (WCE) is accepted as the cost function to optimize the algorithm training process. Suppose that $z_k(x,\theta)$ describes the unnormalized logarithmic probability of pixel x in the k-th category under the given network parameter θ. In that case, the SoftMax function $p_k(x,\theta)$ is defined as:

$$p_k(x,\theta) = \frac{\exp\{z_k(x,\theta)\}}{\sum_{k'}^{K} \exp\{z_{k'}(x,\theta)\}} \quad (26)$$

In (27), K represents the total number of image categories. During forecasting, once equation (26) reaches the maximum, pixel x will be labeled as the k-th category, namely $k* = \arg\max\{P_k(x,\theta)\}$. A semantic segmentation task needs to sum the pixel data loss in each input mini-batch (Li et al., 2018; Guo et al., 2019). Suppose that N denotes the total number of

pixels in the training batch of image data, y_i refers to the real semantic annotation of the pixel x_i , and $p_k(x_i, \theta)$ describes the forecasted probability of pixel x_i belonging to the k-th semantic category, that is, the log-normalized probability, abbreviated as p_{ik} . In that case, the training process aims to find the optimal network parameter $\theta*$ by minimizing the WCE loss function $\ell(x, \theta)$, denoted as $\theta* = \min_{\theta} \ell(x, \theta)$. Training samples with unbalanced categories in brain images usually make the network emphasize some easily distinguishable categories, resulting in poor recognition on some more difficult samples. In this regard, the Online Hard Example Mining (OHEM) strategy (Wen et al., 2019) is adopted to optimize the network training process. The improved loss function is:

$$\ell(x, \theta) = -\frac{1}{\sum_{i=1}^{N}\sum_{k=1}^{K}\delta\left(y_i = k, p_{ik} < \eta\right)} \sum_{i=1}^{N}\sum_{k=1}^{K}\delta\left(y_i = k, p_{ik} < \eta\right)\log p_{ik} \quad (27)$$

In (27), $\eta \in (0, 1]$ refers to the predefined threshold, and $\delta(\cdot)$ describes the symbolic function, which will value 1 if the condition is met and 0 otherwise. The weighted loss function for brain image fusion is defined as:

$$\ell(x, \theta) = -\sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik} q_{ik} \log p_{ik} \quad (28)$$

In (28), $q_{ik} = q(y_i = k|x_i)$ denotes the true label distribution of the k-th category of the pixel x_i , and w_{ik} refers to the weighting coefficient. The following strategy is employed during training:

$$w_{ik} = \frac{1}{\ln(c + p_{ik})} \quad (29)$$

In (29), c is an additional hyperparameter, set to 1.10 based on experience in the simulation experiments of the present work.

Procedures of the proposed algorithm are illustrated in **Figure 5**.

Start
Input: Unimodal features matrix *Xi*
Output : Unimodal features matrix *Zi* after Improved AlexNet
$Z_i \leftarrow \varnothing$
for $t:\left[1,\cdots,L_i\right], i = 1,2,...,N$ **do**
"Poly" learning rate adjustment:
$init\_lr \times \left(1 - \frac{epoch}{\max\_epoch}\right)^{power}$
$init\_lr \leftarrow 5e^{-4}$
$power \leftarrow 0.9$
Searching for the optimal network parameter θ^* by cross entropy loss function $\ell(x, \theta)$;
$\theta^* \leftarrow \min_{\theta} \ell(x, \theta)$
$\ell(x, \theta) \leftarrow -\sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik} q_{ik} \log p_{ik}$
$q_{ik} \leftarrow q(y_i = k \mid x_i)$
$w_{ik} \leftarrow \frac{1}{\ln(c + p_{ik})}$
$c \leftarrow 1.10$
$z_t \leftarrow \mathrm{Re}lu\left(W_z h_t + b_z\right)$//Fully connected layer
$prediction \leftarrow soft\max\left(W_{sft} z_t + b_{sft}\right)$ // Emotion prediction
$Z_i \leftarrow Z_i \cup z_t$
return Z_i
end

FIGURE 5 | Procedures of the proposed algorithm based on Chi Square Test and improved AlexNet.

Simulation Experiments

MATLAB is adopted in the present work to validate the performance of the proposed algorithm. Image data utilized in the simulation experiments come from Chinese paintings at the Thirteenth National Exhibition of Fines Arts in China; they are divided into a training dataset and a test dataset in 7:3. The ratio of each data type in the two datasets shall be consistent. Hyperparameters of the neural network are set as follows: 0.5 dropout, 300 fully connected layer, 120 iterations, and 40 mini-batch to avoid over-fitting. Some state-of-art algorithms are included for performance comparison, including LSTM (Wen et al., 2019), CNN (Ul Haq et al., 2019), RNN (Zhang et al., 2018), AlexNet (Sajjad and Kwon, 2020), and DNN (Jain et al., 2020). Experimental environment configuration includes software and hardware. Regarding software, the operating system is Linux 64bit, the Python version is 3.6.1, and the development platform is PyCharm. Regarding hardware, the Central Processing Unit (CPU) is Intel Core i7-7700@4.2GHz 8 Cores, the internal memory is Kingston DDR4 2,400 MHz 16G, and the Graphics Processing Unit (GPU) is NVIDIA GeForce 1060 8G.

Performance evaluation metrics include Accuracy, Precision, Recall, and F1 score, calculated through the following equations:

$$Acc = \frac{\sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i}}{l} \quad (30)$$

$$\mathrm{Pr}ecision = \frac{\sum_{i=1}^{l} \frac{TP_i}{TP_i + FP_i}}{l} \quad (31)$$

$$Recall = \frac{\sum_{i=1}^{l} \frac{TP_i}{TP_i + FN_i}}{l} \quad (32)$$

$$F1 = \frac{2\mathrm{Pr}ecision \cdot Recall}{\mathrm{Pr}ecision + Recall} \quad (33)$$

In (30) ∼ (33), *TP* represents the number of positive samples forecasted to be positive, *FP* represents the number of negative samples forecasted to be positive, *FN* represents the number of positive samples forecasted to be negative,

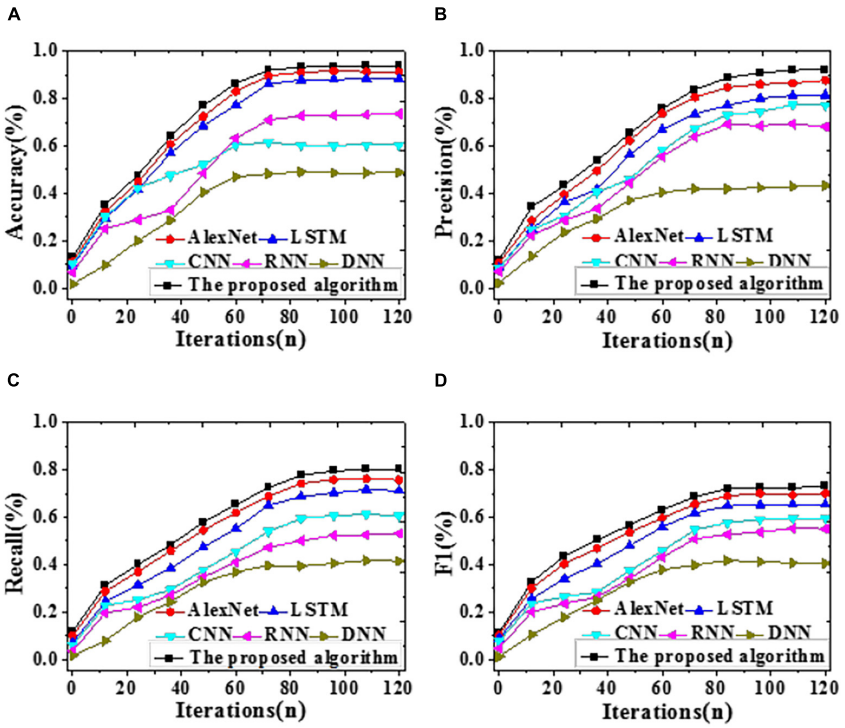


FIGURE 6 | The recognition accuracy of different algorithms with iteration on the training dataset [**(A)** Accuracy; **(B)** Precision; **(C)** Recall; **(D)** F1 score].

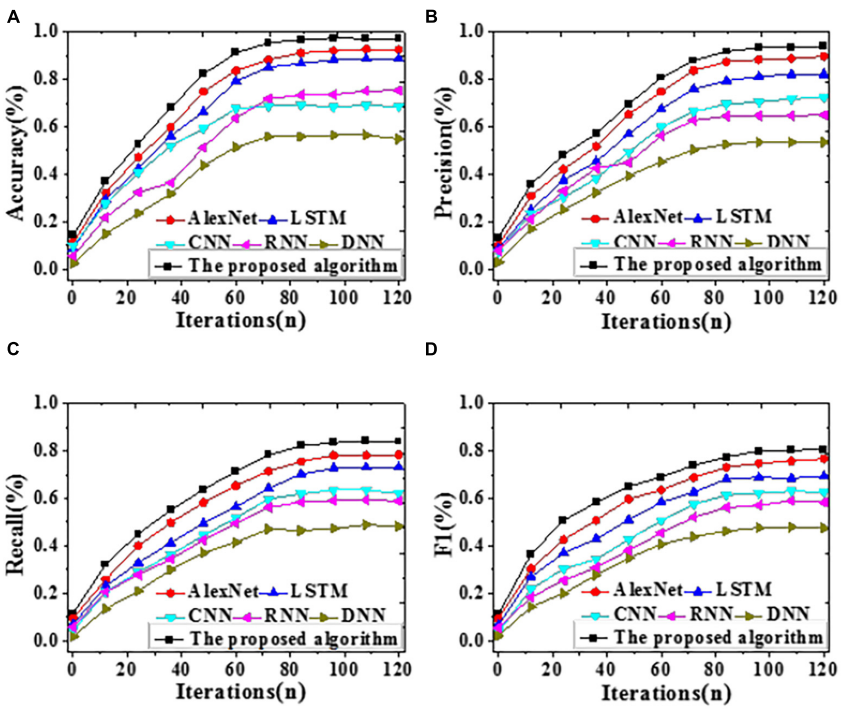


FIGURE 7 | The recognition accuracy of different algorithms with iteration on the test dataset [**(A)** Accuracy; **(B)** Precision; **(C)** Recall; **(D)** F1 score].

and *TN* represents the number of negative samples forecasted to be negative. Accuracy measures the overall classification accuracy, that is, the proportion of samples that forecasted correctly. Recall measures the coverage of positive samples, that is, the proportion of correctly classified positive samples to the total number of positive samples. Precision represents the proportion of examples classified as positive examples to actual positive examples. The most commonly used method is F1 score, which is the weighted harmonic average of precision and recall.

Meanwhile, the emotion recognition efficiency of the algorithm is analyzed from the perspective of SpeedUp. SpeedUp is the ratio of time consumed by the same task running in single processor system and parallel processor system, used to measure the performance and effect of parallel system or program parallelization.

RESULTS AND DISCUSSION

Analysis of Forecasting Performance

To analyze the forecasting performance, the proposed algorithm, LSTM, CNN, RNN, AlexNet, and DNN are included for comparative simulation. **Figures 6**, **7** visualize the results of forecasting accuracy (Accuracy, Precision, Recall, and F1 score). **Tables 1**, **2** illustrate the results of time durations required for training and testing.

As shown in **Figure 6**, on the training dataset, the proposed algorithm can provide an Accuracy of 92.23%, reaching an improvement of at least 4.56% over other algorithms. Remarkably, its Precision, Recall, and F1 score are also the best, at least 3.03% higher than others. To sum up, the proposed algorithm can provide a notably better forecasting performance than DNN, LSTM, AlexNet, RNN, and CNN on the training dataset.

According to **Figure 7**, on the test dataset, the proposed algorithm can provide an Accuracy of 97.11%, reaching an improvement of at least 4.66% over other algorithms. Remarkably, its Precision, Recall, and F1 score are also the best, at least 0.44% higher than others. Hence, the proposed algorithm can provide a remarkably better forecasting performance than DNN, LSTM, AlexNet, RNN, and CNN on the test dataset. **Tables 1**, **2** below illustrate the results of time durations required on the training and test datasets.

Time durations required by all algorithms first decrease sharply then tend to stabilize with epochs; that is, the algorithms converge. In particular, the proposed algorithm requires a training duration of 54.97 s and a testing duration of 23.74 s, remarkably shorter than other algorithms. The proposed algorithm takes less time to make forecasts because of its enhanced generalization ability and accelerated algorithm convergence. In the meantime, the Chi Square Test is specific to emotion recognition, which reduces the time required again. To sum up, the proposed algorithm can achieve higher forecasting accuracy more quickly than other simulated algorithms.

TABLE 1 | Time duration required by different algorithms on the training dataset(s).

	Epochs		
	1.00	**60.00**	**120.00**
The proposed algorithm	90.50	57.91	54.97
AlexNet	90.50	63.44	60.09
LSTM	90.25	67.97	65.87
CNN	91.00	71.49	70.90
RNN	91.51	75.51	75.42
DNN	90.50	80.28	80.20

TABLE 2 | Time duration required by different algorithms on the test dataset(s).

	Epochs		
	1	**60**	**120**
The proposed algorithm	57.81	25.04	23.74
AlexNet	58.15	26.13	24.44
LSTM	57.31	31.03	27.90
CNN	57.65	34.89	29.92
RNN	58.49	39.26	36.81
DNN	59.66	44.64	42.69

Analysis of Acceleration Efficiency

The acceleration efficiency is tested on the training and the test datasets, respectively. Results of time delay error for Chinese painting emotion recognition are visualized in **Figure 8**. The time required and speedup ratio of different algorithms under different data volumes are presented in **Figures 9**, **10**.

As shown in **Figure 8**, errors gradually reduce with iterations on both the training and the test datasets. DNN provides the longest time delay, reaching 455.91 and 356.21 ms, respectively. In contrast, the proposed algorithm provides a time delay approaching zero, the smallest among all simulated algorithms.

According to **Figures 9**, **10**, the proposed algorithm is less sensitive to data growth than other algorithms. Hence, it is suitable to process massive data; the larger the data volume, the higher the speedup ratio, and the greater the acceleration ratio. All algorithms provide slightly better acceleration efficiencies on the test dataset than the training dataset probably because of the absent emotion recognition and analysis path for Chinese paintings during training. Through the adaptive learning, neural networks can process massive amounts of emotion data in Chinese paintings, which remarkably increases the efficiency. To sum up, the proposed algorithm can better recognize and classify human affects in Chinese paintings than other algorithms.

Aiming at the abstract phenomenon of emotion expressed in the field of Chinese painting, the study proposes a multimodal emotion recognition model of Chinese painting based on improved AlexNet neural network and chi square test. The simulation analysis reveals that the recognition and prediction accuracy is significantly higher than that of LSTM, CNN, RNN, AlexNet and DNN model algorithms proposed by scholars in related fields. Among them, the accuracy of the proposed algorithm in the test set reaches 97.11%, and the required time

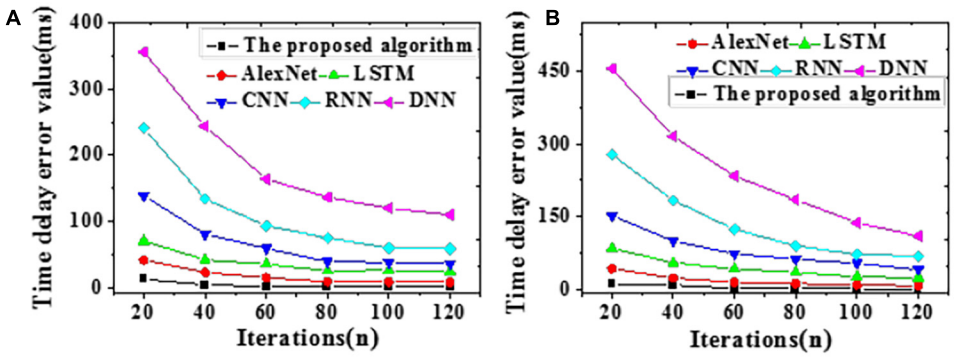


FIGURE 8 | The time delay error of different algorithms for Chinese painting emotion recognition [**(A)** The test dataset; **(B)** The training dataset].

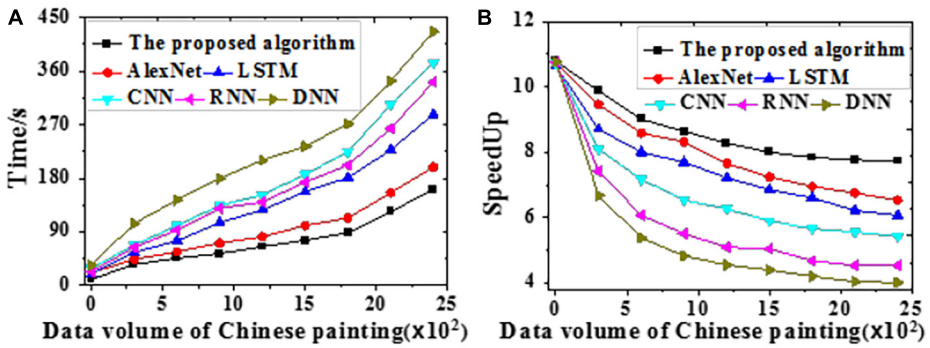


FIGURE 9 | The time required and speedup ratio of different models under different data volumes on the training dataset [**(A)** Time required; **(B)** Speedup ratio].

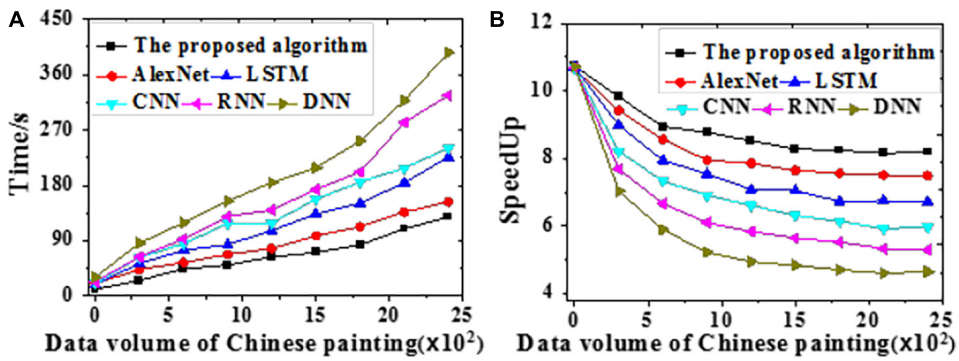


FIGURE 10 | The time required and speedup ratio of different models under different data volumes on the test dataset [**(A)** Time required; **(B)** Speedup ratio].

is only 23.74 s. This may be because the improved AlexNet neural network proposed not only enhances the generalization ability, but also accelerates the convergence rate of the model training process. Moreover, the chi square test is targeted for emotion recognition, which reduces the time required for emotion recognition again. Regarding the recognition efficiency, it is also obvious that the acceleration ratio of the proposed algorithm is higher. This may be because the model has not formed an emotion recognition and analysis path corresponding to Chinese painting in the training process. After autonomous learning in the training process, the neural network can better analyze many emotions of Chinese painting, and the efficiency has been significantly improved. Therefore, the algorithm proposed has a good effect on emotion recognition of Chinese painting.

CONCLUSION

Art such as Chinese painting aims to express the esthetics and emotion of works through visual art elements such as

color, line, and shape. Thus, emotion recognition of images is not only conducive to the management of art information, but also can promote the popularization and promotion of art. To solve "semantic gap" in the emotion recognition task of images such as traditional Chinese painting, this study constructs a multimodal emotion recognition model of Chinese painting based on improved AlexNet neural network and chi square test. The simulation reflects that the emotion recognition accuracy of the improved AlexNet neural network combined with chi square test algorithm model proposed reaches 92.23 and 97.11% in the training set and test set, respectively. The training and test time are stable at about 54.97 and 23.74 s, and the acceleration efficiency is obviously better than other algorithms, which can provide experimental reference for the management, popularization, and promotion of art information in the later stage. However, there are still some deficiencies. First, due to the great differences in the characteristics of traditional Chinese painting in different techniques and content categories, it is difficult to use the same algorithm to identify emotion. In the future, the emotion recognition algorithm for various categories of traditional Chinese painting will be further discussed. Additionally, the mathematical model of traditional Chinese painting emotion will be established, and the traditional Chinese painting emotion will be calculated and processed more accurately to create greater value in the digital management and protection of national quintessence. Second, this study models and analyzes the emotion of images such as Chinese painting in the 13th National Exhibition of Fine Arts. But it is not clear to what extent the proposed method and its theory can be applied to other types of painting and natural images. Therefore, future work will focus on evaluating the effectiveness of this method except for the field of various categories of art.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

JL and DC designed the multi-modal emotion recognition algorithm of Chinese paintings. ZZ and NY conducted the evaluation and analysis of experimental data, compared other algorithms include LSTM, CNN, RNN, AlexNet, and DNN under the guidance of ZL. All authors participated in the process of writing and revising the manuscript.

FUNDING

This work was the concluding outcome of the Key Subject of Shandong Art Science under grant number QN202008253, Shandong Art Education Special Project under grant number ZY20201335, and Shandong Social Science Planning Project under grant number 21CWYJ16.

REFERENCES

Avots, E., Sapiński, T., Bachmann, M., and Kamińska, D. (2019). Audiovisual emotion recognition in wild. *Mach. Vis. Appl.* 30, 975–985. doi: 10.1007/s00138-018-0960-9

Bi, H., Xu, F., Wei, Z., Xue, Y., and Xu, Z. (2019). An active deep learning approach for minimally supervised polsar image classification. *IEEE Trans. Geosci. Remote Sens.* 57, 9378–9395. doi: 10.1109/tgrs.2019.2926434

Cai, H., Qu, Z., Li, Z., Zhang, Y., Hu, X., and Hu, B. (2020). Feature-level fusion approaches based on multimodal EEG data for depression recognition. *Inf. Fusion* 59, 127–138. doi: 10.1016/j.inffus.2020.01.008

Chen, M., Jiang, Y., Cao, Y., and Zomaya, A. Y. (2020). CreativeBioMan: a Brain- and Body-Wearable, Computing-Based, Creative Gaming System. *IEEE Syst. Man Cybern. Mag.* 6, 14–22. doi: 10.1109/msmc.2019.2929312

Ćosić, K., Popović, S., Šarlija, M., Mijić, I., Kokot, M., Kesedžić, I., et al. (2019). New tools and methods in selection of air traffic controllers based on multimodal psychophysiological measurements. *IEEE Access* 7, 174873–174888. doi: 10.1109/access.2019.2957357

Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., et al. (2019). Ce-net: context encoder network for 2d medical image segmentation. *IEEE Trans. Med. Imaging* 38, 2281–2292. doi: 10.1109/tmi.2019.2903562

Guo, Z., Li, X., Huang, H., Guo, N., and Li, Q. (2019). Deep learning-based image segmentation on multimodal medical imaging. *IEEE Trans. Radiat. Plasma Med. Sci.* 3, 162–169.

Haberl, M. G., Christopher, C., Lucas, T., Daniela, B., Sébastien, P., Bushong, E. A., et al. (2018). Cdeep3m—plug-and-play cloud-based deep learning for image segmentation. *Nat. Methods* 15, 677–680. doi: 10.1038/s41592-018-0106-z

Hammad, M., Liu, Y., and Wang, K. (2018). Multimodal biometric authentication systems using convolution neural network based on different level fusion of ECG and fingerprint. *IEEE Access* 7, 26527–26542. doi: 10.1109/access.2018.2886573

He, X., and Zhang, W. (2018). Emotion recognition by assisted learning with convolutional neural networks. *Neurocomputing* 291, 187–194. doi: 10.1016/j.neucom.2018.02.073

Hwang, S., Hong, K., Son, G., and Byun, H. (2020). Learning CNN features from DE features for EEG-based emotion recognition. *Pattern Anal. Appl.* 23, 1323–1335. doi: 10.1007/s10044-019-00860-w

Jain, D. K., Zhang, Z., and Huang, K. (2020). Multi angle optimal pattern-based deep learning for automatic facial expression recognition. *Pattern Recognit. Lett.* 139, 157–165. doi: 10.1016/j.patrec.2017.06.025

Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., and Zareapoor, M. (2018). Hybrid deep neural networks for face emotion recognition. *Pattern Recognit. Lett.* 115, 101–106. doi: 10.1016/j.patrec.2018.04.010

Jia, H., Peng, X., Song, W., Lang, C., and Sun, K. (2019). Multiverse optimization algorithm based on lévy flight improvement for multithreshold color image segmentation. *IEEE Access* 7, 32805–32844. doi: 10.1109/access.2019.2903345

Kim, W., Kanezaki, A., and Tanaka, M. (2020). Unsupervised learning of image segmentation based on differentiable feature clustering. *IEEE Trans. Image Process.* 29, 8055–8068. doi: 10.1109/tip.2020.3011269

Li, L., Fu, H., and Tai, C. L. (2018). Fast sketch segmentation and labeling with deep learning. *IEEE Comput. Graph. Appl.* 39, 38–51. doi: 10.1109/mcg.2018.2884192

Lin, Z., Chen, Y., Ghamisi, P., and Benediktsson, J. A. (2018). Generative adversarial networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* 56, 5046–5063.

Liu, S., Yang, J., Agaian, S. S., and Yuan, C. (2021). Novel features for art movement classification of portrait paintings. *Image Vis. Comput.* 108:104121. doi: 10.1016/j.imavis.2021.104121

Nalepa, J., Myller, M., and Kawulok, M. (2019). Validating hyperspectral image segmentation. *IEEE Geosci. Remote Sens. Lett.* 16, 1264–1268. doi: 10.1109/lgrs.2019.2895697

Sajjad, M., and Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access* 8, 79861–79875. doi: 10.1109/access.2020.2990405

Spoorthi, G. E., Gorthi, S., and Gorthi, R. (2018). Phasenet: a deep convolutional neural network for two-dimensional phase unwrapping. *IEEE Signal Process. Lett.* 26, 54–58. doi: 10.1109/lsp.2018.2879184

Sukumarana, A., and Mb, A. (2019). A Brief Review of Conventional and Deep Learning Approaches in Facial Emotion Recognition. *Artificial Intelligence for Internet of Things* 101.

Ul Haq, I., Ullah, A., Muhammad, K., Lee, M. Y., and Baik, S. W. (2019). Personalized movie summarization using deep cnn-assisted facial expression recognition. *Complexity* 2019, 1–10. doi: 10.1155/2019/3581419

Walia, G. S., and Rohilla, R. (2021). A Contemporary Survey of Multimodal Presentation Attack Detection Techniques: challenges and Opportunities. *SN Comput. Sci.* 2:49.

Wang, G., Li, W., Zuluaga, M. A., Pratt, R., Patel, P. A., Aertsen, M., et al. (2017). Interactive medical image segmentation using deep learning with image-specific fine-tuning. *IEEE Trans. Med. Imaging* 37, 1562–1573. doi: 10.1109/tmi.2018.2791721

Wang, Z., Lian, J., Song, C., Zhang, Z., Zheng, W., Yue, S., et al. (2019). SAS: painting Detection and Recognition via Smart Art System With Mobile Devices. *IEEE Access* 7, 135563–135572. doi: 10.1109/access.2019.2941239

Wei, J., Yang, X., and Dong, Y. (2021). User-generated video emotion recognition based on key frames. *Multimed. Tools Appl.* 80, 14343–14361. doi: 10.1007/s11042-020-10203-1

Wen, S., Wei, H., Yang, Y., Guo, Z., Zeng, Z., Huang, T., et al. (2019). Memristive LSTM network for sentiment analysis. *IEEE Trans. Syst. Man Cybern. Syst.* 51, 1794–1804.

Xu, Y., and Li, Q. (2021). Music Classification and Detection of Location Factors of Feature Words in Complex Noise Environment. *Complexity* 2021, 1–12. doi: 10.1155/2021/5518967

Yang, D., Ye, X., and Guo, B. (2021). Application of Multitask Joint Sparse Representation Algorithm in Chinese Painting Image Classification. *Complexity* 2021, 1–11. doi: 10.1155/2021/5546338

Yuan, J., Feng, Z., Dong, D., Meng, X., Meng, J., and Kong, D. (2020). Research on Multimodal Perceptual Navigational Virtual and Real Fusion Intelligent Experiment Equipment and Algorithm. *IEEE Access* 8, 43375–43390. doi: 10.1109/access.2020.2978089

Zhang, J., Miao, Y., Zhang, J., and Yu, J. (2020). Inkthetics: a Comprehensive Computational Model for Aesthetic Evaluation of Chinese Ink Paintings. *IEEE Access* 8, 225857–225871. doi: 10.1109/access.2020.3044573

Zhang, T., Zheng, W., Cui, Z., Zong, Y., and Li, Y. (2018). Spatial–temporal recurrent neural network for emotion recognition. *IEEE Trans. Cybern.* 49, 839–847. doi: 10.1109/tcyb.2017.2788081

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.