Research on Learners' Emotion Recognition Method in Teaching Environment

Minghan Wang
Tianjin University of Technology, Tianji 300382, China.

Abstract: In this study, the authors propose a method that combines CNN and LSTM networks to recognize facial expressions. To handle illumination changes and preserve edge information in the image, the method uses two different preprocessing techniques. The preprocessed image is then fed into two independent CNN layers for feature extraction. The extracted features are then fused with an LSTM layer to capture the temporal dynamics of facial expressions. To evaluate the method's performance, the authors use the FER2013 dataset, which contains over 35,000 facial images with seven different expressions. To ensure a balanced distribution of the expressions in the training and testing sets, a mixing matrix is generated. The models in FER on the FER2013 dataset with an accuracy of 73.72%. The use of Focal loss, a variant of cross-entropy loss, improves the model's performance, especially in handling class imbalance. Overall, the proposed method demonstrates strong generalization ability and robustness to variations in illumination and facial expressions. It has the potential to be applied in various real-world applications such as emotion recognition in virtual assistants, driver monitoring systems, and mental health diagnosis.

Keywords: Emotion Recognition; Teaching Environment; Facial Expressions

Introduction

Facial expressions on the human face are a fundamental way for humans to convey their inner emotions. Human beings understand each other's inner feelings by observing the changes in their facial expressions, and at the same time express their inner emotions intuitively and accurately through facial expressions. The use of facial expression recognition systems in real life has broad application prospects, and it can simply and quickly obtain customers' recognition of service attitudes and service quality. For example, the catering industry can measure the deliciousness of dishes based on the user's facial expressions and emotions. Medical institutions can analyze the patient's state through the patient's facial expression and emotion, so as to provide more detailed medical care. With the rapid development of face recognition technology, facial expression recognition has become one of the main research directions of academic experts. The research content covers various disciplines, including psychology., biological neurology, information science, etc., is a complex subject with great challenges and high research value. Facial expression recognition technology includes six parts, involving the input of facial expression images, the detection of human faces, and the recognition of facial expression features. With the in-depth study, it has driven the development of artificial intelligence, computer vision, biometric features and psychology, and promoted the proposal of new methods and theories .

It mainly applies convolutional neural network, and has made remarkable achievements in facial expression recognition. Since the neural network is trained through the dataset, its performance is usually better than traditional methods. However, when real-time is considered, it is difficult for deep learning to process temporal and spatial signals to obtain better emotion recognition performance. Many two-dimensional CNNs cannot recognize temporal information. In this case, researchers have developed a method that combines temporal and spatial features, namely CNN-LSTM. Inspired by this framework, this paper employs a CNN-LSTM model for image sequences acquired from datasets and real-time environments.

1. Literature Review

AlexNet was a significant breakthrough in deep learning and computer vision, as it demonstrated that deep convolutional neural networks could achieve superior performance in image classification tasks. Since its introduction, numerous variations of AlexNet have been proposed, and its architecture has inspired the design of many other successful deep neural networks, such as VGGNet, GoogleNet, and ResNet.

These networks have shown remarkable performance on large-scale datasets and have enabled the development of many practical applications, such as self-driving cars, medical diagnosis, and surveillance systems.

When using CNN network to extract features, time information will be ignored, and it will be inaccurate when processing real-time video information. Therefore, scholars have developed several deep learning networks that can simultaneously learn temporal and spatial features (Shahabinejad, et al., 2021 & Akhand, et al ., 2021) . The neural network model combining CNN and LSTM is used to jointly learn spatial and temporal features to complete different target recognition tasks.

Due to the continuous development of deep learning technology, expression analysis technology is gradually divided into traditional analysis methods and machine learning analysis methods (Altaher, et al., 2020) . On the one hand, traditional analysis methods use linear discriminant analysis method. LDA is the most typical tool for feature extraction and data dimensionality reduction in the field of pattern recognition. It is easier to classify. ICA is a way to describe multivariate data sets using a linear coordinate system. It belongs to the unsupervised, adaptive component analysis method, ICA algorithm The main part is the optimization criterion, which is used to judge whether the result is good or bad. The combination of ICA and LDA is to map the collected n-dimensional information to n-dimensional. After the goal of dimensionality reduction has been achieved, the image information is described twice, and then the expression is extracted. Key information (Borgalli, et al., 2020) . On the other hand, the local binary model LBP (Local inary Pattem) uses the amplitude and the image frame as a product, and then uses LBP to input the product to the support vector machine to extract the expression information. Through the layer-by-layer feature extraction and layer-by-layer abstraction between the established networks, the final information is purified to obtain the expression information in the input image frame. Purpose (Mellouk, et al., 2021) .

Contrastive loss, triplet loss, center loss, and Focal loss allow samples to be evenly distributed around the center of the class to minimize intra-class differences, but the calculation efficiency is too low. By focusing on the parameter √, the model can pay more attention to difficult-to-classify samples and improve the classification performance of the model, but it cannot solve the labeling sample problem (Nonis, et al., 2021 & Ab Wahab, et al., 2021) . Facial expression recognition using deep learning techniques has become a popular research topic in recent years. To improve the performance of facial expression recognition models, various loss functions have been proposed to optimize the model parameters. Among them, Contrastive loss, triplet loss, center loss, and Focal loss are commonly used in deep learning-based facial expression recognition models.

Contrastive loss and triplet loss are designed to minimize the distance. Center loss, on the other hand, focuses on minimizing intra-class variance by forcing samples to be evenly distributed around the center of the class. Focal loss is designed to address the problem of class imbalance by giving more weight to hard-to-classify samples. However, these loss functions can be computationally expensive and may not be efficient for real-time applications. To address this issue, researchers have proposed various techniques to improve the efficiency of these loss functions, such as using the parameter √ to focus on difficult-to-classify samples.

In this paper, the researchers have used the Focal loss function in the CNN-LSTM model for facial expression recognition. The Focal loss function is designed to focus on hard-to-classify samples and has been shown to be effective in addressing class imbalance problems. By using the Focal loss function, the researchers aim to improve the classification performance of the CNN-LSTM model for facial expression recognition. Overall, the development and application of efficient loss functions for facial expression recognition models are crucial to improve their performance and enable their use in real-world applications.

2. Research methods

2.1 Image preprocessing

Facial key point detection is an important part of facial analysis, and it is the basic technology for applications such as face recognition, expression judgment, and 3D facial remodeling. The key point is to reflect the facial features of various parts, and with the

development of technology and precision With the improvement of the level, the number of face key points has also increased from the original five points to more than two hundred points today. Due to the increasing development of various technologies in computer vision, face key point detection can be divided into Parametric method and non-parametric method The way of parametric model can be divided into local-based method and global-based method according to its external modeling, using partial or global appearance features for modeling, and predicting key points through graphical models. At present, the best and most widely used method is the method based on deep learning.

Key point processing is based on deep learning.

In order to unify the input data, first detect the face area, then crop and adjust the detected face size to 96×96 pixels, the adjusted probability density function calculation formula is:

$$P(G_k) = \frac{N_k}{N} \quad (1)$$

Among them, Gk is the drawing image, N is the number of occurrences of Gk, and N is the total number of pixels in the image.

Next, the quality of the image is improved through the transformation function T( ), and the weighted histogram is equalized as:

$$P_w(G_k) = T\{P(G_k)\}$$
$$P_w(G_k) = \begin{cases} P_w(G_k), P(G_k) > \tau_1 \\ 0, P(G_k) < \tau_2 \end{cases} \quad (2)$$

Among them, $\tau 1$ and $\tau 2$ are the upper and lower limits, defined as follows, where β is the weight function, and $\beta<1$ and non-zero.

Therefore, the calculation formula of histogram equalization is as follows:

$$P_w(G_k) = \left[\frac{P(G_k) - \tau_2}{\tau_1 - \tau_2} \right] \times \tau, \tau_1 < \tau < \tau_2 \quad (3)$$

Use the above formula to process images with different lighting conditions as one of the inputs to the CNN. Second, edge enhancement is performed on the same group of images, and edge pixels are enhanced using Euclidean distance and chamfering distance. The distance is to use the nearest distance value to mark the pixel value in the image, and the Euclidean distance uses the average distance.

2.2 Feature extraction

In the network used in this paper, the batch adjustment and ReLU activation function are used, so the network can effectively learn the characteristics of the image. The specific formula is as follows:

$$\mathbf{ReLU}(x) = \mathbf{max}(0, x) \quad (3)$$

The CNN used in this article consists of 5 convolutional layers, the filter size is 5×5, and the pooling layer selects the maximum pooling. After 5 feature extractions, the extracted features are sent to the fully connected layer, and integrated with the LSTM layer. Together. The specific parameters of each layer are shown in Table 1.

2.3 Focal loss

In the expression recognition task, the loss function in the multi-classification task will lead to poor classification performance due to the imbalance of samples. Therefore, in response to this problem, the standard cross entropy loss function is improved, and a focal loss function (Focal loss) is proposed. The formula is as follows:

$$FL = -\alpha(1-p_i)^{\gamma} \ln p_i \quad (3)$$

If the sample is misclassified, p_i is an infinitesimal number, so $(1 - p)^{\gamma}$ is close to 1, and when the correct classification is pi is 1,

$(1 - p)^{\gamma}$ is close to 0. Focalloss can be applied to difficult samples by controlling the modulation coefficient, and achieve the purpose of balancing the number of samples of different categories through the balance parameter α.

3. Results

3.1 Experimental environment

The computer processor used in this experiment is AMD Ryzen 7-1700 model, the graphics card is GTX 1080 TITAN, and the memory is 16 GB. The neural network part is built using the open source Keras module, and the software programming environment is python3 0.

3.2 Facial Expression Dataset

In order to prove the effectiveness of the proposed architecture, experiments were done on the facial expression database of FER2013. The number of pictures of the six expressions in each database is shown in Table 1.

3.3 Analysis of results

1) Use two methods of histogram equalization and edge enhancement to preprocess FER2013, and use the model proposed in this paper to verify the accuracy of the processed pictures. The accuracy of the two preprocessing models is compared as shown in Table 3. See The accuracy rate obtained by using edge enhancement is 3%~8% higher, and the accuracy rate is 80.14% in the FER2013 data set after using the edge enhancement preprocessing method .

2) In order to verify the effectiveness on the cross dataset, a mixing matrix was made according to the FER2013 dataset.

3) Finally, compare with the accuracy of several existing international methods on the FER2013 dataset. The accuracy rate of the CNN-LSTM model proposed in this paper is 9.65% higher than the current algorithm with the highest accuracy rate. The specific data are shown in Table 4.

Discussion

Facial expression recognition is a challenging task due to various factors, such as changes in lighting conditions, facial expressions, and occlusions. In this paper, the researchers propose a new method for facial expression recognition that addresses these challenges by using edge enhancement and histogram equalization as preprocessing methods. The edge enhancement technique is used to improve the contrast and sharpness of the facial features, while histogram equalization is used to normalize the distribution of pixel intensities in the image. These preprocessing techniques help to deal with various lighting conditions in different environments.

In addition, the researchers use a combination of two modules, CNN and LSTM, to learn both spatial and temporal features from the facial images. The CNN module is used to extract spatial features from the images, while the LSTM module is used to capture the temporal dynamics across the frames. To optimize the model parameters, the researchers use the Focal loss function, which is designed to address the problem of class imbalance by giving more weight to hard-to-classify samples. The Focal loss function has been shown to be effective in improving the classification performance of deep learning models. The method achieves better classification performance when the training samples are relatively small relative to the local information. However, there is still a need for further research to improve the robustness of the model in more complex environments, such as outdoor settings. In future research, it is expected that the proposed method can be further optimized and extended to be applied, such as emotion recognition in human-robot interaction, healthcare, and education.

Facial expression recognition technology has become an important research area. The technology involves three main steps: face detection and collection, expression feature extraction, and facial expression analysis. The first step of face detection and collection is critical in facial expression recognition, as it involves locating the face in an image or video and collecting the necessary data for expression analysis. In recent years, face detection and collection have become increasingly mature as an independent research direction, with numerous algorithms and techniques being developed for this purpose.

The second step is the core issue of facial expression recognition research, as it involves extracting effective features from the face that can describe the differences between different expressions. There are various techniques for extracting expression features, including geometric features, appearance features, and texture features. These features can be extracted from different regions of the face, such as the eyes, mouth, and nose. The third step of facial expression analysis involves selecting the correct facial expression

recognition classifier for the extracted expression features to identify the expression. The choice of classifier depends on the specific requirements of the application, such as real-time performance, accuracy, and robustness.

Improving the accuracy of facial expression classification is a relevant scientific research goal, as it has important applications in various fields, such as human-robot interaction, healthcare, and education. Deep learning techniques have shown remarkable performance in facial expression recognition, and researchers are exploring new methods to improve the accuracy and efficiency of these techniques. The use of convolutional neural networks (CNNs) in deep learning for facial key point detection and expression recognition has shown significant progress in recent years. This approach involves training a CNN to locate the key points of the face, such as the eyes, nose, and mouth, and then combining this information with the CNN to estimate facial expressions.

While this approach has shown promising results in accurately estimating facial expressions, it still faces challenges in specific uncontrollable scenes, such as image occlusion or background interference factors. These challenges can lead to misjudgment due to changes in the front and rear frames of the image or too many background factors.

Another challenge of this approach is its limited real-time performance due to the use of CNNs with many layers and two input and output streams. Although the estimation accuracy is high, the real-time performance of the system is average, which can limit its use in real-world applications. To address these challenges, researchers are exploring ways to balance the real-time performance. This includes exploring the use of lightweight CNN models that can achieve high accuracy with fewer layers and parameters. Additionally, techniques such as data augmentation, transfer learning, and optimization algorithms can be used to improve the efficiency.

In summary, while the use of CNNs for facial key point detection and expression recognition has shown significant progress, there are still challenges to be addressed to achieve robust performance in various real-world scenarios. Continued research and development in this area can lead to the development of more efficient and accurate facial expression recognition systems.

References

[1] Zahara, L., Musa, P., Wibowo, EP, Karim, I., & Musa, SB (2020, November). The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi. In *2020 Fifth international conference on informatics and computing (ICIC)* (pp. 1-9). IEEE.

[2] Akhand, MAH, Roy, S., Siddique, N., Kamal, MAS, & Shimamura, T. (2021). Facial emotion recognition using transfer learning in the deep CNN. Electronics, *10* (*9*), 1036.

[3] Shahabinejad, M., Wang, Y., Yu, Y., Tang, J., & Li, J. (2021, December). Toward personalized emotion recognition: A face recognition based attention method for facial emotion recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)* (pp. 1-5). IEEE.

[4] Borgalli, MRA, & Surve, S. (2022, March). Deep learning for facial emotion recognition using custom CNN architecture. In *Journal of Physics: Conference Series* (Vol. 2236, No. 1, p. 012004). IOP Publishing.

[5] Altaher, A., Salekshahrezaee, Z., Abdollah Zadeh, A., Rafieipour, H., & Altaher, A. (2020). Using multi-inception CNN for face emotion recognition. *Journal of Bioengineering Research, 3* (1), 1-12.

[6] Nonis, F., Barbiero, P., Cirrincione, G., Olivetti, EC, Marcolin, F., & Vezzetti, E. (2021). Understanding abstraction in deep CNN: an application on facial emotion recognition. *Progresses in Artificial Intelligence and Neural Systems,* 281-290.

[7] Ab Wahab, MN, Nazir, A., Ren, ATZ, Noor, MHM, Akbar, MF, & Mohamed, ASA (2021). Efficientnet -lite and hybrid CNN-KNN implementation for facial expression recognition on raspberry pi. *IEEE Access, 9,* 134065-134080.

[8] Mellouk, W., & Handouzi, W. (2020). Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science, 175,* 689-694.

Tables and figures

Table 1 Number of 6 expressions in FER2013

data set	anger	disgust	fear	happy	sad	surprise
FER2013	4 953	5 47	5 121	8 989	6 077	4 002

Table 2 Accuracy % using different preprocessing methods

Model	Histogram equalization ratio	edge enhancement
FER2013 _	7 7.32	**8 0.14**

Table 3 FER2013 mixing matrix %

expression	anger	disgust	fear	happy	sad	surprise
anger	**9 1.3**	2 .6	0.9 _	0	0	0
disgust	2.7	**90.1**	1.2	0	0	0
fear	0	0	**88.2**	4.8	1.35	5.6
happy	1.1	0	0.9	**92.6**	0	0
sad	0	0	0.99	0	**91.1**	0
surprise	0.3	0	0	0	0.7	**89.2**

Table 4 Accuracy % of recognition by different methods on the FER2013 dataset

Model	**Accuracy**
CNN+Improve_Softmax	70.91
DNNRL	70.60
SHCNN	69.10
VGG+Focal loss	72.49
LSTM+CNN	**82.14**

Appendix:

Soft skills: Through the preliminary analysis of the thesis topic, I have improved in data collection and analysis, code writing, result analysis, and paper structure layout. This process made me clear about the important value of the paper. The first is logic, the overall logic of the paper can help me and readers improve the reading ability of the article; Secondly, the level of analysis, I made it clear in the processing of experimental results that in addition to the description of the phenomenon, it is also necessary to strengthen the analysis of the essential phenomenon behind the content. Finally, this process improves the comprehensive ability of thesis writing.

Hard skills: Construction of CNN networks, construction of LSTM networks, Keras, sklearn, numpy, pandas, etc Commonly used function calls in Python and Origin to draw experimental diagrams