DOI: 10.53469/wjimt.2024.07(05).07

Facial Expression Recognition Based on the FER2013 Dataset

Yang Lou, Dan Li

School of Computer and Software, Jincheng College, Sichuan University, Chengdu 610000, Sichuan, China

Abstract: ***Facial emotions are a way to express one's thoughts and also an effective way to understand the emotions of others. Nowadays, with the rapid development of technology, computers can also recognize facial expressions through convolutional neural networks, deep learning, and other methods, and classify the results. Throughout the entire experiment, we chose FER2013 data as the training set for our model, which ultimately achieved an accuracy of around 62%. We also compared it with the SFEW dataset. The emergence of facial expression recognition will increase in the future, and its application in teaching supervision is what we are exploring here. Its main function can be used for invigilation, attendance, checking class status, and so on.***

Keywords: Facial expression recognition; FER2013 dataset; Convolutional neural network; Covariance; Teaching supervision.

1. THE SIGNIFICANCE OF FACIAL EXPRESSION RECOGNITION

1.1 Significance

Nowadays, human-computer interaction systems are ubiquitous in daily life, and it is precisely for this reason that the intelligence and reliability of human-computer interaction systems are becoming increasingly important[1-3]. The use of facial expression recognition in human-computer interaction systems can greatly enhance their intelligence. By recognizing users' facial expressions and identifying their emotional changes, it can make communication between products and users more natural [4-7]. Of course, facial expression recognition is not only used in human-computer interaction systems, as technology advances and facial expression recognition tends to be optimized [8]. Its application scope is constantly expanding, such as safe driving, medical pain detection, teaching supervision, and so on.

1.2 Current Status of Facial Expression Recognition

Facial expression recognition technology has been developed for many years, and its main core components are face detection, facial feature extraction, and expression classification. For face detection, it is no longer a difficult problem. Currently, the main focus is on how to extract facial features. Regarding the classification of facial expressions, they are now divided into basic expressions and complex expressions. Basic expressions are divided into 7 types: happy, sad, surprised, scared, angry, annoying, and calm. Basic facial expressions cannot represent the complexity of human emotions in daily life, while the facial motion coding system FACS can be used to represent more complex facial expressions [9-11].

Because in daily life, we often express basic expressions, there is currently little research on complex expressions, and many studies are specifically focused on the seven basic expressions. In addition, obtaining the dataset is relatively difficult because research on facial expression recognition requires a large amount of facial expression data for training, and the current dataset is limited. SFEW, CK+, FER2013 and other datasets are some of the ones we frequently use [12-15].

Facial expression recognition technology first started abroad and has rapidly developed in the 21st century, dividing it into static recognition of images or photos and dynamic recognition of videos and recordings. At the beginning, people used traditional methods to extract facial features, such as LBP features, principal component analysis (PCA), and Gabor wavelets [16]. And with the emergence of deep learning, using deep learning to extract features has become the most commonly used method. The earliest research on facial expression analysis in our country was introduced by Harbin Institute of Technology in 1997. After continuous research and development, we have reached a good level in facial recognition. Wu, Z. (2024), introduces the Meta-Path Guided Attention Aggregation Network (MPAAGN), which integrates meta-paths, attention mechanisms, and GraphSAGE for

efficient node classification in heterogeneous networks, making it applicable to any large-scale graph data requiring advanced relationship modeling and node classification [17].

2. TECHNICAL CORE

2.1 Convolutional Neural Networks

Convolutional Neural Network (CNN) is a type of feedforward neural network, as shown in Figure 1. In addition to the input-output layers, other layers play a similar role in CNN as the importance of the heart to humans. In addition to processing such files, convolutional neural networks can also handle information such as audio and text. As long as the data can be converted into image format, it can be processed by convolutional neural networks.

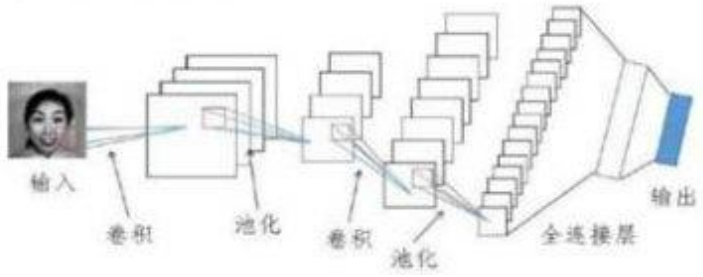


Figure 1: Basic Structure of Convolutional Neural Network

2.1.1 Convolutional Layer

Using a certain part of an image for matching to avoid using the whole is called convolution. Convolution operation is the process of comparing different parts of two images, which are called convolution kernels and also known as features. And the convolution kernel needs to be manually set at the beginning, and after a lot of calculations, an optimal convolution kernel will be obtained. The operation process is shown in Figure 21, with the input matrix on the left and the convolution kernel in the middle. It will perform dot product operation on the left side with a certain step size to obtain the output matrix on the right side. This is the most basic operation of the convolution layer.

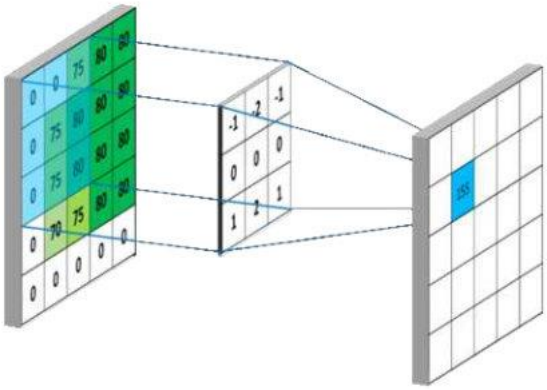


Figure 2: Convolution operation process

2.1.2 Pooling layer

Pooling is the process of reducing the output of a convolutional layer based on the convolutional kernel, in order to achieve the goal of reducing computational complexity. Although it may sacrifice some information, it is only acceptable. Pooling generally includes max pooling and average pooling. As shown in Figure 3, max pooling is the process of extracting the maximum number within a specified range of 2 * 2 boxes, which is 22, while average pooling is the process of taking the average of those four numbers and then extracting them, resulting in a final value of 9. Max pooling is commonly used.

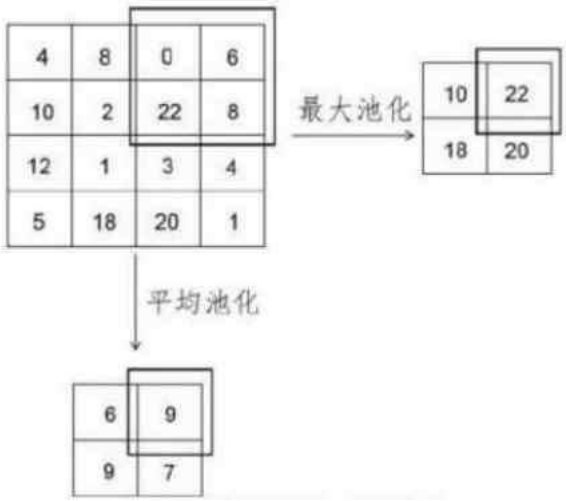


Figure 3: Maximum pooling and average pooling

2.1.3 Fully Connected Layer

The fully connected layer is similar to a classifier in that it can classify them based on the extracted features in convolutional neural networks. The task of the fully connected layer is to collect the filtered images from higher levels, convert these feature information into votes, that is, weights. The values of each matrix need to be connected to neurons, and each neuron needs to multiply its own weight. The total number of votes obtained is the highest output type. A fully connected layer can have more than one layer, and with each additional fully connected layer, the judgments made by the convolutional neural network will become more accurate.

2.1.4 ReLu function

The ReLu function also plays a significant role in convolutional neural networks, as it is an activation function that primarily introduces nonlinearity in computation. Its mathematical principle is to convert negative values less than 0 in the features into 0, because in matrix operations, multiplying 0 with other values can facilitate calculation.

2.2 Differences between Parties

Covariance difference pooling refers to changing the pooling steps in the pooling layer from selecting the maximum value to extracting covariance. The features extracted from covariance are more accurate compared to other methods.

2.2.1 Covariance

If you want to measure two random variables, you need to use covariance. The formula is shown in Figure 4. If the calculated result is positive, it indicates a positive correlation between these two random variables; If the result is negative, it indicates a negative correlation between the two; If it is zero, it indicates that the two are not correlated, which is known as mutual independence in correlation statistics.

$$\begin{aligned} Cov(X,Y) &= E[(X-E[X])(Y-E[Y])] \\ &= E[XY]-2E[Y]E[X]+E[X]E[Y] \\ &= E[XY]-E[X]E[Y] \end{aligned}$$

Figure 4: Covariance formula

2.2.2 Covariance Matrix

When performing image expression recognition, the covariance matrix is obtained by flattening the output of the convolutional layer and performing vector operations; In video expression recognition, the output of the fully connected layer is used as the image set feature and then calculated from it.

3. DATASETS

The dataset cannot be omitted in any model, as it directly affects the final result of the entire model. In model algorithms, we usually divide the dataset into training set, validation set, and testing set. As the name suggests, the training set is used to train the model; The validation set is used to adjust the learning model; The test set is used to evaluate the overall performance of the model.

3.1 Common Datasets

In the field of facial recognition, there are already many related datasets available for experimentation and comparison, such as CK+facial expression database, JAFFE expression dataset, FER2013 expression dataset, etc. These datasets can be found online, which can help our team model achieve higher accuracy.

3.2 FER2013 Dataset

The dataset contains over 35000 grayscale images, divided into 7 different facial expressions. The 7 expressions are represented by numerical labels ranging from 0 to 7, where 0 represents anger, 1 represents disgust, 2 represents fear, 3 represents happiness, 4 represents sadness, 5 represents surprise, and 6 represents nature. The size of each image is 48 * 48 pixels, but the dataset is saved in a CSV file and the images are not directly provided. As shown in Figure 5, where emotion, pixels, and usage represent the attributes of the data, the emotion column represents the category of facial expressions, the pixels column represents the specific data of the image, and the usage column represents which dataset it belongs to. For example, 0 represents anger, and the ending English represents for use in the training set.

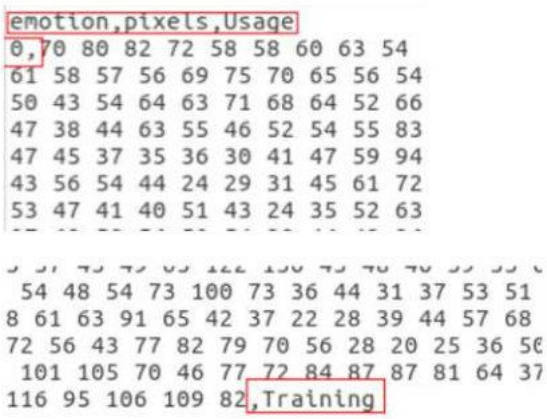


Figure 5: FER2013 dataset

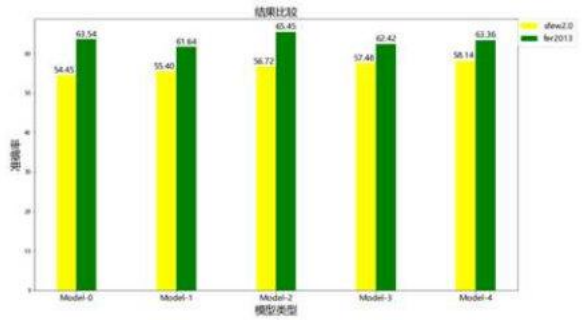
3.3 SFEW2.0 Dataset

The SFEW2.0 dataset is a dataset used in a facial expression recognition competition abroad, which is often used for static facial expression recognition, that is, image-based expression recognition. The vast majority of images in this dataset are from characters in movies, and different expressions were cropped and classified to create this dataset. The SFEW2.0 dataset is also divided into 7 types of facial expressions.

3.4 Comparison of Results

During testing, the original dataset, SFEW2.0 dataset, had an accuracy of only about 58%, while after replacement, the model's accuracy fluctuated around 62%. The main reason is that the SFEW2.0 dataset comes from movie screenshots, which contain a lot of irrelevant information and have a certain impact on detection. Additionally, many facial expressions come from the night, making face detection a bit difficult and resulting in lower accuracy. However, the FER2013 dataset does not have as much irrelevant information and is only grayscale images without the influence of light. However, there are also some incorrect labels in the FER2013 dataset, so its accuracy has not reached a very high level. Through experimental comparison, it was found that covariance pooling can indeed improve the accuracy of facial expression recognition. In addition, by adjusting the number of iterations, the

number of fully connected layers, and changing the model framework, these methods can help us improve the accuracy of the experiment.



编号	迭代次数	全连接层层数	准确率
1	5000	3	62.46
2	5500	3	64.82
3	5500	5	65.34

Figure 6: Accuracy of SFEW dataset and FER2013 dataset under different models and results under different parameters

4. APPLIED TO TEACHING SUPERVISION

Due to the continuous development of technology, face-to-face teaching methods have also been somewhat impacted, and teaching modes have become increasingly diverse, such as the emergence of online platform teaching. However, the emergence of new developments and problems, as well as new educational methods, has also given rise to many issues. Some students play with their phones and sleep in class, but the teacher cannot immediately notice. But facial expression recognition can help teachers detect such situations in a timely manner. Facial recognition systems can also play many other roles, such as preventing cheating in exams, improving attendance rates, and so on.

4.1 Preventing Cheating in Exams

In middle school, there are usually two invigilators during exams, and there are only forty to fifty candidates in one examination room, so cheating is not serious. In universities, there are often nearly a hundred candidates in a test room, but there are still only two invigilators, which provides an "opportunity" for candidates who do not study hard in their studies. Many people find someone to take the exam on their behalf in order to get a good score, but there are still some loopholes through checking student ID cards, ID cards, and other means. However, through facial expression recognition, we can verify student information one by one and reject proxy exam students. By examining the expressions of candidates during the exam, we can also promptly detect whether they have engaged in cheating behavior, greatly ensuring the fairness of the exam.

4.2 Increase Attendance Rate

No matter what university it is, there will always be situations where students seek help from others to substitute for classes, and this phenomenon is not uncommon. By using facial expression recognition, not only can this phenomenon be eliminated, but it can also save the time for teachers to call names. After analyzing the expressions of classmates during class, we can determine whether they have deviated or not, thereby improving the quality of teaching.

5. SUMMARY

Nowadays, by learning convolutional neural networks and accumulating self expansion, we can easily achieve this project. It is not only applicable in the field of teaching supervision, but also widely used in various fields such as healthcare, safe driving, and human-computer interaction. In this facial expression recognition used for teaching supervision, we achieved higher accuracy by changing the dataset and adjusting parameters, but the final accuracy was still relatively low. So, in the future, we plan to improve accuracy by replacing data and adding more facial expression features to increase our ability to recognize complex facial expressions.

REFERENCES

[1] Wu, Z., Wang, X., Huang, S., Yang, H., & Ma, D. (2024). Research on Prediction Recommendation System Based on Improved Markov Model. Advances in Computer, Signals and Systems, 8(5), 87-97.
[2] Wu, Z. (2024). MPGAAN: Effective and Efficient Heterogeneous Information Network Classification. Journal of Computer Science and Technology Studies, 6(4), 08-16.
[3] Zhou Benjun Research on Facial Expression Recognition Based on Convolutional Neural Networks [D] Nanjing: Nanjing University of Posts and Telecommunications, 2019.
[4] Gao Wen, Jin Hui Analysis and Recognition of Facial Expression Images [J] Journal of Computer Science, 1997, 20 (9): 782-789.
[5] Jiang, L., Yu, C., Wu, Z., & Wang, Y. (2024). Advanced AI framework for enhanced detection and assessment of abdominal trauma: Integrating 3D segmentation with 2D CNN and RNN models. arXiv preprint arXiv:2407.16165.
[6] Yan, H., Wang, Z., Xu, Z., Wang, Z., Wu, Z., & Lyu, R. (2024). Research on image super-resolution reconstruction mechanism based on convolutional neural network. arXiv preprint arXiv:2407.13211.
[7] Ji, H., Xu, X., Su, G., Wang, J., & Wang, Y. (2024). Utilizing Machine Learning for Precise Audience Targeting in Data Science and Targeted Advertising. Academic Journal of Science and Technology, 9(2), 215-220.
[8] https://www.kaggle.com/c/challenges-in-representation-learn-ing-facial-expression-recognition-challenge/data.
[9] Xu, S., Wu, P., & Chen, Y. (2024). Interfacial thermal conductance in 2D WS2/MoSe2 and MoS2/MoSe2 lateral heterostructures. Computational Materials Science, 245, 113282.
[10] Wu, P., Iquebal, A. S., & Ankit, K. (2023). Emulating microstructural evolution during spinodal decomposition using a tensor decomposed convolutional and recurrent neural network. Computational Materials Science, 224, 112187.
[11] Wang, W., & Osaragi, T. (2024). Lognormal distribution of daily travel time and a utility model for its emergence. Transportation research part A: policy and practice, 183, 104058.
[12] Peng, Q., Ding, Z., Lyu, L., Sun, L., & Chen, C. (2022). RAIN: regularization on input and network for black-box domain adaptation. arXiv preprint arXiv:2208.10531.
[13] Chen, H., Yang, Y., & Shao, C. (2021). Multi-task learning for data-efficient spatiotemporal modeling of tool surface progression in ultrasonic metal welding. Journal of Manufacturing Systems, 58, 306-315.
[14] Cao, Y., Cao, P., Chen, H., Kochendorfer, K. M., Trotter, A. B., Galanter, W. L., ... & Iyer, R. K. (2022). Predicting ICU admissions for hospitalized COVID-19 patients with a factor graph-based model. In Multimodal AI in healthcare: A paradigm shift in health intelligence (pp. 245-256). Cham: Springer International Publishing.
[15] Zheng Ren, "Balancing role contributions: a novel approach for role-oriented dialogue summarization," Proc. SPIE 13259, International Conference on Automation Control, Algorithm, and Intelligent Bionics (ACAIB 2024), 1325920 (4 September 2024); https://doi.org/10.1117/12.3039616
[16] Z. Ren, "Enhancing Seq2Seq Models for Role-Oriented Dialogue Summary Generation Through Adaptive Feature Weighting and Dynamic Statistical Conditioninge," 2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE), Guangzhou, China, 2024, pp. 497-501, doi: 10.1109/CISCE62493.2024.10653360.
[17] Wu, Z. (2024). MPGAAN: Effective and Efficient Heterogeneous Information Network Classification. Journal of Computer Science and Technology Studies, 6(4), 08-16.