Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/95/20241661

Occluded facial expression recognition based on deep learning

Xuetong Zhao

Sichuan Agricultural University, Ya'an City, Sichuan Province, 625014, China

17318215867@163.com

Abstract. When employing Convolutional Neural Networks (CNNs) for facial expression recognition, several challenges are often encountered, such as facial occlusions, the limited size of reliable expression datasets, and inadequate precision in recognition outcomes. This paper preprocesses the dataset to enhance its reliability. By leveraging data synthesis and augmentation techniques, it employs a method of randomly generating occlusion blocks to integrate and expand the dataset. Based on the ResNet-18 network, the model is optimized by incorporating an attention mechanism, thereby improving the network's precision and robustness in recognizing facial expressions.

Keywords: Deep Learning, Attention Mechanism, Occluded Face.

1. Introduction

Facial expressions are one of the important ways for humans to express emotions and intentions. By recognizing and understanding expressions, computers can more accurately capture the emotional state of users, thereby enhancing the naturalness and efficiency of human-computer interaction. This provides a powerful tool for the fields of computer vision and artificial intelligence to deeply understand human emotions. It not only promotes the development of human-computer interaction technology, enabling machines to respond and adapt to human emotions more naturally, but also shows a wide range of application potential in various fields such as mental health assessment, security monitoring, education, and market research.

Since the 18th century, many scholars have begun researching facial expression recognition. At the end of the 19th century, Darwin wrote the book "The Expression of the Emotions in Man and Animals," elaborating on the characteristics of facial expressions in humans and animals. In 1971, American psychologists Friesen and Ekman et al., after analysis, systematically classified basic facial expressions into six categories: happiness, anger, sadness, surprise, disgust, and fear [1], which is also the current recognized form of expression classification. Early research on facial expression recognition focused on rule-based methods, where researchers used manually designed feature extraction and simple classifiers to identify basic expressions. With the development of machine learning technology, traditional machine learning methods such as Support Vector Machines (SVM) [2] and Random Forests [3] began to be introduced into expression recognition research. These methods improved recognition accuracy by learning low-level image features, such as Gabor filter responses [4] and Local Binary Patterns (LBP) [5]. In recent years, the breakthrough progress of deep learning technology has greatly promoted the development of the field of facial expression recognition. Convolutional Neural Networks (CNNs) have become mainstream due to their powerful feature learning capabilities. Researchers have begun to train

DOI: 10.54254/2755-2721/95/20241661

deep neural networks with large annotated datasets to achieve higher-level learning of expression features. Krizhevsky et al.'s [6] AlexNet model achieved success in the ImageNet competition, laying the foundation for the widespread application of deep learning in computer vision. Simonyan et al. [7] proposed several VGG network variants with different depths (16 and 19 layers), which uniformly use small-sized 3x3 convolutional filters, stacked by multiple convolutional layers (with ReLU activation functions) and max-pooling layers, making it easy to explore the impact of network depth on performance.

With the development of computer technology, extracting and comparing facial features using computers has become a hot research direction. The data analysis by computers makes the recognition of expressions more rational and well-founded. In recent years, thanks to the self-learning ability of CNNs, more and more research has been conducted on expression recognition using Convolutional Neural Networks, but there are still issues such as the lack of reliable datasets, facial occlusions in non-controlled environments, and the need for improved precision in recognition results. This paper will analyze the above issues and seek optimization strategies.

2. Datasets

2.1. CK+

The CK+ dataset [8], also known as the Extended Cohn-Kanade dataset, is a widely used public dataset for facial expression recognition research. It provides a controlled experimental environment, allowing researchers to evaluate and compare different facial expression recognition algorithms under standardized conditions. The CK+ dataset has been expanded from the Cohn-Kanade dataset [9], not only labeling the categories of expressions but also including detailed annotations of Action Units. The CK+ dataset includes 123 subjects of varying ages, genders, and ethnicities, with a total of 593 video sequences, each recording the transition from a neutral expression to a specific expression. Among them, 327 sequences are labeled with emotional tags, with emotion categories including 0=Neutral, 1=Anger, 2=Contempt, 3=Disgust, 4=Fear, 5=Happiness, 6=Sadness, and 7=Surprise.

The image sequences of the CK+ dataset are typically stored in the .h5 file format, which contains image pixel data and corresponding label data. When processing the dataset, it is necessary to store the image sequences according to different emotional categories and ensure that the training set does not include image sequences from the test set. The structure of the CK+ dataset is designed to meet the various needs of facial expression recognition research. In addition to the basic image sequences, the dataset also provides facial landmark annotations and FACS coding, which are crucial for understanding the mechanisms of expression generation and for automated facial feature extraction.

Figure 1. A sample graph of seven expressions in the CK+ dataset

2.2. FER2013

Images in the FER2013 dataset are primarily collected from the internet and encompass subjects of various ages, genders, and ethnicities. To ensure consistency and standardization of the images, preprocessing such as cropping and resizing is performed. The data is divided into training, validation, and testing sets, with proportions of 80%, 10%, and 10%, respectively, for the widespread use in training and testing various facial expression recognition algorithms.

The FER2013 dataset consists of approximately 30,000 facial expression images, with a uniform image size of 48×48 pixels. The dataset includes 7 basic expressions. The number of images for each expression type is roughly equivalent, but the number of images for the disgust expression is the least, with only 600 images, while the other expression categories have nearly 5,000 images each.

DOI: 10.54254/2755-2721/95/20241661

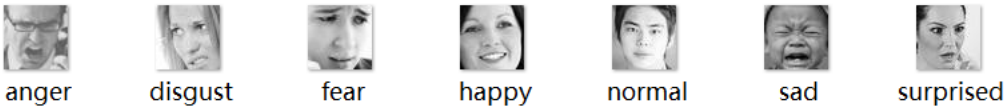

Figure 2. A sample graph of seven expressions in the FER2013 dataset

2.3. Dataset Processing

Since the FER2013 dataset is typically provided in CSV file format, where the facial expression data exists in the form of pixel values, conversion is required to obtain the images. The pandas library in Python is used to read the CSV file, converting each pixel value string into a 48x48 two-dimensional array. Subsequently, the OpenCV library is used to save the NumPy array as an image file, and the images are categorized and stored in different folders according to the expression types.

In the real world, faces may be occluded by objects such as hands, hair, glasses, and masks. To simulate the situation of occluded faces in real-world scenarios, a custom occlusion transformation is created using torchvision.transforms. The method involves generating random coordinates for the occlusion area to create a rectangular or circular region, which is then superimposed onto the original image. By simulating occlusions [10], the diversity of the training data can be increased, preventing the model from overfitting to a specific data distribution.

3. Local Attention Mechanism

Occlusions can interfere with the regular process of facial feature extraction, necessitating the design of a feature extraction method that is robust against occlusions. Moreover, this study is based on the improvement of the ResNet-18 network [11], which, as a shallow deep neural network, has limited feature extraction capabilities, especially when processing complex images, where there may be issues with information loss or confusion. By introducing an attention mechanism, specifically the channel and spatial attention of the CBAM module [12], the network can more selectively focus on important feature channels and spatial locations, thereby enhancing the quality and diversity of feature representation. The improvement in the model's learning ability helps the model generalize better between the training and testing sets, thus performing more stably and reliably in real scenarios. At the same time, the CBAM module does not significantly increase the complexity and computational cost of the network, especially in lightweight network structures like ResNet-18, where it can be effectively integrated and applied.

The combination of spatial and channel attention can more comprehensively consider the importance of overall features, thereby improving the effect of facial feature extraction. The channel attention module compresses and describes each channel's feature map by introducing two global pooling operations (global average pooling and global max pooling). Subsequently, a fully connected layer is used to generate the weight for each channel, i.e., the channel attention weights. The channel attention weights are normalized through the sigmoid function and then multiplied with the original feature maps to enhance meaningful channels and suppress irrelevant ones, making the network more focused on channels that contribute to expression classification. The spatial attention mechanism can identify important areas in the image and enhance the feature expression of these areas. It combines the average and maximum values of the feature maps and uses convolutional layers to learn the importance weights of spatial positions. When processing occluded facial images, spatial attention can help the model better locate and utilize visible facial features, reducing the interference of occluded areas.

4. Model Optimization

The core idea of ResNet is the introduction of residual blocks, which solve the problem of gradient vanishing and gradient explosion in deep networks through skip connections. This is particularly important for occluded face recognition because occlusions may lead to missing or incomplete information in certain areas of the input image. Residual learning helps the network to better adapt to such situations, improving the accuracy and robustness of recognition. ResNet-18 is a relatively shallow deep neural network model, with sufficient depth to extract complex features, yet lighter than some

DOI: 10.54254/2755-2721/95/20241661

deeper models (such as ResNet-50 or larger). This balance allows ResNet-18 to maintain good performance in handling occluded facial expression recognition tasks without being overly complex, which could make the training and inference processes too costly.

To further enhance the ability of feature extraction, a local attention mechanism is introduced. After the output of each residual block, a CBAM module is added to enhance the feature representation. The output of the CBAM module will replace the original output of the residual block, serving as the input for the next residual block. After all residual blocks have been processed, the feature maps are subjected to a global feature extraction through an average pooling layer, followed by one or more fully connected layers for the final prediction of the classification task.

Table 1. Comparison with experimental results of different networks

Network	CK+/%	FER2013/%
AlexNet[13]	94.17	68.87
MIANet[14]	95.76	72.28
CNN[15]	84.37	65.00
APRNET50[16]	94.95	73.00
VGG16[17]	90.65	60.40
ours	96.12	74.86

5. Conclusion

This paper aims to improve the accuracy of recognizing facial expressions under occlusion by modifying the dataset images and optimizing the model network structure. Firstly, the images in the CK+ and FER2013 datasets are standardized and unified to meet the input requirements of the model. Data augmentation and occlusion simulation operations are added, artificially adding occlusions to the existing facial expression datasets to simulate real-world occlusion scenarios, and a small amount of random noise is added to the images to enhance the model's robustness to input perturbations. Based on the ResNet-18 network as the framework, a local attention mechanism is introduced, with a CBAM module added after each residual block to provide a richer feature representation for the output of each residual block, thereby enhancing the model's ability to recognize occluded facial expressions. The methods presented in this paper have demonstrated higher accuracy on the two datasets while maintaining the computational complexity of the model. At the same time, the model in this paper can be further improved, and there is still room for improvement in terms of accuracy.

References

[1] Ekman, P., & Friesen, W. V. (1978). Facial action coding system: a technique for the measurement of facial movement.

[2] Zuo, K. L. (2004). A Study of Automatic Facial Expression Analysis and Recognition S ystem [Doctoral Dissertation, Tianjin University]. https://kns.cnki.net/kcms2/article/abstract?v=kxD1c6RDvBwteB7ZYhNkeGzvnsW8jFJiLPCm5nXxHyHH45Qgafkytykg5peKKTNF5XfBm3Cj__WTcvy47K90yvomE6UvjP37XC4qd0fG_t3esi0eFbNf2ECQKc6YBgqZT4knXLAQt8irVNliTcu3R1kFpwvlWMzv7tmVNo4aDRugx3XtPE7ivIib6ZNeBu_P&uniplatform=NZKPT&language=CHS

[3] Walecki, R., Rudovic, O., Pavlovic, V. (2015, 4-8 May 2015). Variable-state latent conditional random fields for facial expression recognition and action unit detection. 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG),

[4] He, L., Zou, C., Bao, Y. (2005). The research advance of facial expression recognition. Acta Electronica Sinica, (01), 70-75.

[5] Wang, T., Peng, X., Zhu, J. (2021). Research on smiley face recognition algorithm based on fusion of geometric features and LBP features. Electronic Test, (23), 52-54. doi:10.16520/j.cnki.1000-8519.2021.23.016.

DOI: 10.54254/2755-2721/95/20241661

[6] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Commun. ACM, 60(6), 84–90. https://doi.org/10.1145/3065386

[7] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556.

[8] Lucey, P., Cohn, J. F., Kanade, T. (2010, 13-18 June 2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops,

[9] Kanade, T., Cohn, J. F., & Yingli, T. (2000, 28-30 March 2000). Comprehensive database for facial expression analysis. Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580),

[10] Nan, Y. & Hua, Q. (2022). Research progress of deep learning methods for occlusion facial expression recognition. Computer Applications and Software, (02), 321-330. doi:10.19734/j.issn.1001-3695.2021.08.0307.

[11] Wang, J., Yang, Y., & He, Y. (2018). Pornographic Images Recognition Framework Based on Multi-Classification and ResNet. Computer Systems Applications, (09), 100-106. doi:10.15888/j.cnki.csa.006517.

[12] Woo, S., Park, J., Lee, J.-Y. (2018). CBAM: Convolutional Block Attention Module. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss, Computer Vision – ECCV 2018 Cham.

[13] Zhang, X. (2022). Research and Application of Facial Expression Recognition Based on Deep Learning [Master's Thesis, Chongqing University of Posts and Telecommunications]. https://link.cnki.net/doi/10.27675/d.cnki.gcydx.2022.000803

[14] Luo, S., Li, M., & Chen, M. (2023). Multi-Scale Integrated Attention Mechanism for Facial Expression Recognition Network. Computer Engineering and Applications, (01), 199-206.

[15] Agrawal, A., & Mittal, N. (2020). Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. The Visual Computer, 36(2), 405-412. https://doi.org/10.1007/s00371-019-01630-9

[16] Chen, J., & Xu, Y. (2022). Expression Recognition Based on Convolution Residual Network of Attention Pyramid. Computer Engineering and Applications, (22), 123-131.

[17] Guo, X., Shen, Z., & Wang, X. (2023). Face expression recognition based on improved VGG network. Journal of Changchun University of Technology, (01), 52-57. doi:10.15923/j.cnki.cn22-1382/t.2023.1.08.