Faceknow: Facial Expression Recognition by a Global-Local Network with a Sub-Images-Related Contextual Attention Mechanism

Jing Li¹, Yonggang He², Zhiwei Li¹, Yuda Luo¹, Xiaoyun Zhang³ and Hongbo Ni¹

¹School of Computer Science, Northwestern Polytechnical University, Xi’an, China
²School of Software, Northwestern Polytechnical University, Xi’an, China
³DeRUCCI Health Sleep Co., Ltd. Shenzhen Branch, Shenzhen, China

Abstract. Recognizing health status through human faces is a challenging research topic. Among them, facial recognition expressions can indirectly reflect the inner health status, which has very significant commercial and research value. Most research work on facial expression recognition uses traditional methods, and the accuracy of traditional methods highly depends on feature extraction. Deep learning has already promoted the research on facial expression recognition. This paper proposes a dual-branch network that uses global facial information and local information obtained by using the attention mechanism to merge and identify human facial emotional information. Use of shared pre-training module to extract low-level semantic information of global and also local images. The dual-branch network architecture utilizes the attention module to capture the relationship between different sub-images to fuse the local features of the face. Experimental results demonstrate that the accuracy of the CK+ Dataset reaches 95.96%, which is improved compared to other existing methods.

Keywords. Facial expression recognition; deep learning; attention mechanism; convolutional neural network; image classification.

1. Introduction
Facial expression is the easiest way for humans to express emotions, and it can take account of the current psychological state of humans to a certain extent. The generally recognized types have 6-8 core expressions, such as happiness, anger, sadness, surprise, disgust, and fear [1]. Facial expression recognition has a wide range of applications in areas such as autonomous driving, human-computer interaction, and medical treatment [2]. Recently, many people have conducted research on facial expression recognition, but it is challenging to be able to recognize expression categories with high accuracy.

Facial expression recognition is a sub-field of image recognition, which studies the expression of human emotions on the face. Humans use facial expressions to convey information about their inner feelings. This way of expression is objective and can be measured. Through facial expression recognition technology, we can better understand human feelings. Facial expression recognition usually has two ways to extract features, traditional feature extraction, and deep learning feature extraction. Most traditional methods use manual features, such as local binary pattern (LBP) [3], non-negative matrix factorization (NMF) [4], and sparse learning [5] for facial expression recognition [6]. Recently, deep learning has also improved the accuracy of facial expression recognition. Convolutional neural network is often used in computer vision, and has achieved great success in this field [7]. The convolutional neural network can adaptively extract the features of each part, and adapt them to various
classification problems after adjustment. However, features extracted using CNN are usually for the global facial expression image, ignoring local information. Changes in facial expressions usually appear near the eyes, nose, and mouth. These local information is very important for facial expression recognition.

The attention mechanism was originally used in the field of machine translation, and it has now been used in computer vision. The attention mechanism can be explained intuitively by using the human visual mechanism. By learning from the intermediate attention map, and then applying the product of each element on the attention map and the source feature map to weight the importance of different features, the most useful features are selected for classification. Using the attention mechanism can improve the weight of the key features then improve the accuracy of facial expression recognition [2].

We propose a two-branch CNN structure, and contextual attention related to a sub-images module is used in this structure. In the dual-branch structure, one branch is to extract the features of the global image, and the other branch is to extract the local features by the sub-images-related contextual attention mechanism module, and finally, the two branches are fused for classification. Embedding the attention mechanism module into the network can make the CNN pay more attention to important features, which are conducive to the improvement of the accuracy of recognition.

The main contributions for this paper are listed below:

- We use a two-branch convolutional neural network structure, one branch extracts global information, and one branch adds a sub-images-related contextual attention mechanism module to extract local information. It can ignore areas that are not related to facial expression recognition, while sub-images-related contextual attention mechanisms can take into account the relationship between different sub-images during feature fusion, and improve the accuracy of recognition.

The Shared pre-training module is used to extract low-level features of global and local images, such as edge information and angle information. It can reduce the complexity of the model and expands the data set.

We conducted experiments in the CK+ dataset, the most used dataset in the field of facial expression recognition. The performance of our model is compared with other advanced expression recognition algorithm models, and our model performs better in the experiment.

The following are the details of other parts of the article. In Section 2, we will briefly introduce the existing work and algorithms of facial expression recognition. In Section 3, we will introduce our proposed method in detail. In Section 4, we will show that the experiment and result analysis on the data set. We will give the conclusion in Section 5.

2. Related Work
We divide the traditional facial expression recognition algorithms into two categories. Algorithms based on extracting geometric features, for example, Active Appearance Model (AAM) [9] and facial action unit AU [10], etc. Algorithms based on extracting texture features, for example, local binary pattern (LBP) [10] and Gabor wavelet [11]. The development of deep learning has further promoted the development of facial expression recognition. This method uses deep learning models to automatically extract features and classifications. Commonly used methods include convolutional neural network (CNN) and recurrent neural network (RNN). More scholars have begun to develop more complex and diverse the deep learning model is used for the research of facial expression recognition. Reference [12] designed a new type of convolutional neural network structure, using a convolution kernel to extract hidden features. And using maximum pooling to reduce the dimensionality of the extracted hidden features. [13] proposed an end-to-end architecture called spatial-temporal convolution feature of nested LSTM. The structure can learn the appearance features of facial expressions and time-related features at the same time.

An end-to-end structure is proposed in [14]. This structure uses the attention module, so the classification performance is significantly improved. In [15], a deep learning model based on weakly supervised patches is proposed for joint intensity estimation of multiple AUs, and the attention mechanism is further enhanced through a learnable task-related context.
3. Proposed Method

3.1. Network Structure
We used a two-branch convolutional neural network structure and added a sub-images-related contextual attention mechanism module to the branch of the partial image.

As shown in figure 1, the facial expression recognition algorithm is mainly divided into three parts: preprocessing, feature extraction, and classification. Specifically, we first crop the facial area of the person in the image through the facial cropping algorithm, and then divide it into multiple sub-images, and put the cropped overall image and sub-images into a model with shared weights. This model is constructed using the first 14 layers of pre-trained GoogleNet. This structure is used to extract low-level features from the global image and local images. For the features extracted from the segmented sub-images, we add an attention module to it. And combined with a learnable context to enhance the attention mechanism, used to capture the relationship between different sub-images. The network structure of this deep learning algorithm is shown in figure 2.

3.2. Shared Pre-Training Module
The shared pre-training module uses the first 14 layers of GoogleNet for pre-training so that it can extract low-level semantic features of the face. Using the shared pre-training module to extract global and local features enables the model to learn common low-level features, such as edge features and angle features. This reduces the complexity of the model and expands the data set. After inputting the image into the sharing and training module, it is input into the last 8-layer model as a high-level semantic information extraction module.
3.3. Sub-Images-Related Contextual Attention Module

After we segment the image directly, the attention value of each part should be different, and there should be a certain connection between the sub-images of different parts. For example, if a person has a happy expression, then the corners of his mouth will be upward, the end of the eyes will also rise, and wrinkles will appear in certain areas of the face. These positions are distributed in different sub-images, and we need to capture the correlation between different sub-images to improve the accuracy of model recognition. Therefore, we set up a series of context vectors related to sub-images to represent the connection between our different sub-images.

During preprocessing, we divide the image into N sub-images, \( I = \{I_1, I_2, ..., I_N\} \) represents a collection of N sub-images, \( F = \{F_1, F_2, ..., F_N\} \) represents low-level semantic features extracted from N sub-images, and \( F' = \{F'_1, F'_2, ..., F'_N\} \) represents high-level semantic features extracted from N sub-images.

The high-level semantic features extracted by the deep learning model of the later layers can be expressed as:

\[
F'_n = f(I_n; \Theta_f), \quad n \in \{1, 2, ..., N\}
\]

where \( f \) refers to the convolutional neural network, where \( \Theta_f \) is the parameter in the network.

The high-level semantic features extracted after adding the attention module related to the sub-image can be expressed as:

\[
F' = \sum_{n=1}^{N} \alpha(I_n; I, C_n)F'_n
\]

Among them, \( \alpha \) represents the attention value, \( I_n \) represents a certain sub-image, \( I \) represents the overall image, \( C_n \) represents the context of the sub-image whose sequence number is \( n \), and \( F'_n \) represents the extracted high-level feature.

\[
\alpha(I_n; I, C_n) = \frac{\exp(w^T\tanh(w_cC_n+w_hF'_n))}{\sum_j \exp(w^T\tanh(w_cC_n+w_hF'_j))}
\]

The softmax function is employed to normalize it so as to ensure that the sum of all attention weights is 1. Among them, \( w, w_c \) and \( w_h \) are the parameters that can be learned. The calculation of this attention value involves all sub-images and the context related to other sub-images. By adding an attention module, the local features of the N sub-images are fused together.

3.4. Fusion and Classification

Considering that the features of the global image and the local image have complementary features, we fuse the global feature with the local feature. Fusion is divided into two parts, feature-level, and decision-level. Feature-level fusion is the local feature fusion in Section 3.3. At the decision level, we consider the global and local outputs and then merge them to generate a score. This score represents the final classification result:

\[
O = \beta P_{\text{global}} + (1 - \beta) P_{\text{local}}
\]

where \( P_{\text{global}} \) represents the output of the global image, \( P_{\text{local}} \) represents the output of the local image, \( \beta \) is a balance parameter, we set it to 0.5 based on experience, which represents the output category.

4. Experiment

4.1. CK+ Dataset

The CK+ Dataset [16] contains 593 facial expression sequences of 123 subjects, of which 327 sequences have expression tags. These expression sequences are labeled into seven categories. We select three frames near the peak frame in each sequence as training samples and use 10-fold cross-validation.
4.2. Parameter Setting
The Adam algorithm is used to optimize the parameters, and a batch size of 128 is used to train 40 epochs. Set the learning rate to 0.0001. When the loss on the verification set no longer changes, adjust the learning rate to 0.00001 and continue to train. The weight attenuation is set to 0.0005. The data set is divided into 10 parts, 9 parts of which are used as training data, and 1 part is used as test data in turn. And the average of the accuracy of the results of 10 times is used as the accurate estimation of the final algorithm.

4.3. Experimental Results
Table 1 compares the performance of our method with the latest method of expression recognition using the CK+ Dataset. Our method uses global images and local images. STRNN only considers the global image as input for analysis and may ignore the local face information. In addition, STRNN requires multiple sequence frames as input, which requires relatively high input conditions, and the calculation of sequence frames is more complicated. IB-CNN uses facial action unit labels as input, and the training samples are very limited. Moreover, the spatial relationship between different action units is not analyzed, so the accuracy of our method is better than that of IB-CNN.

| Method          | Accuracy Rate |
|-----------------|---------------|
| MSR [17]        | 91.4%         |
| 3DCNN-DAP [18]  | 92.4%         |
| Inception [19]  | 93.2%         |
| IB-CNN [20]     | 95.1%         |
| STRNN [21]      | 95.4%         |
| **The proposed algorithm** | **95.96%** |

5. Conclusion
In this article, we propose a two-branch network to extract global and local features for facial expression recognition. A shared pre-training module is used to extract low-level semantic information about global and local images. We set up a sub-images-related contextual attention mechanism module for the extraction of local information to capture the relationship of the sub-images so as to strengthen the fusion of local features. After experimental comparison, our method shows a significant improvement over other methods.

Acknowledgments
This work was supported by the Key Research and Development Program in Shaanxi Province (grant no. 2020ZDLGY04-08).

References
[1] Jeong D, Kim B G and Dong S Y 2020 Deep joint spatiotemporal network (DJSTN) for efficient facial expression recognition Sensors 20 (7) 1936.
[2] Li J, Jin K, Zhou D, Kubota N and Ju Z 2020 Attention mechanism-based cnn for facial expression recognition Neurocomputing 411 340-350.
[3] Shan C, Gong S and Mcowan P W 2009 Facial expression recognition based on local binary patterns: a comprehensive study Image and Vision Computing 27 (6) 803-16.
[4] Zhi R, Flierl M, Ruan Q and Kleijn W B 2011 Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition IEEE Trans Syst Man Cybern B Cybern 41 (1) 38-52.
[5] Lin Z, Liu Q, Peng Y, Bo L and Metaxas D N 2012 Learning active facial patches for expression analysis Computer Vision and Pattern Recognition (CVPR) 2012 IEEE Conference on IEEE pp 2562-69.

[6] Shan L and Deng W 2018 Deep facial expression recognition: A survey IEEE Transactions on Affective Computing 99.

[7] Krizhevsky A, Sutskever I and Hinton G 2012 Imagenet classification with deep convolutional neural networks Advances in Neural Information Processing Systems 1097-105.

[8] Cootes T F, Edwards G J and Taylor C J 2001 Active Appearance Models IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) 681-85.

[9] Alkawaz M H, Mohamad D, Basori A H, et al. 2015 Blend shape interpolation and FACS for realistic avatar 3D Research 6 (1) 38.

[10] Shan C, Gong S and Mcowan P W 2009 Facial expression recognition based on local binary patterns: A comprehensive study Image and Vision Computing 27 (6) 803-16.

[11] Zhang S, Li L and Zhao Z 2012 Facial expression recognition based on Gabor wavelets and sparse representation IEEE International Conference on Signal Processing pp 816-19.

[12] Chen X, Yang X, Wang M and Zou J 2017 Convolutional neural network for automatic facial expression recognition International Conference on Applied System Innovation (ICASI) pp 814-17.

[13] Yu Z, Liu G, Liu Q and Deng J 2018 Spatio-temporal convolutional features with nested LSTM for facial expression recognition Neurocomputing 317 50-57.

[14] Fernandez P, Peña F A G, Ren T I and Cunha A 2019 FERAtt: Facial expression recognition with attention net IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp 837-846.

[15] Zhang Y, et al. 2019 Context-aware feature and label fusion for facial action unit intensity estimation with partially labeled data IEEE/CVF International Conference on Computer Vision (ICCV) pp 733-742.

[16] Lucey P, Cohn J F, Kanade T, et al. 2010 The extended Cohn-Kanade dataset (CK)+: A complete dataset for action unit and emotion-specified expression IEEE Conf. Comput. Vis. Pattern Recognit. Workshops pp 94-101.

[17] Rifai S, Bengio Y, Courville A, Vincent P and Mirza M 2012 Disentangling factors of variation for facial expression recognition European Conference on Computer Vision (Berlin: Springer) pp 808-22.

[18] Liu M, Li S, Shan S, Wang R and Chen X 2014 Deeply learning deformable facial action parts model for dynamic expression analysis Asian Conference on Computer Vision (Springer) pp 143-57.

[19] Mollahosseini A, Chan D and Mahoor M H 2016 Going deeper in facial expression recognition using deep neural networks IEEE Winter Conference on Applications of Computer Vision (WACV) pp 1-10.

[20] Han S, Meng Z, Khan A S and Yan T 2017 Incremental boosting convolutional neural network for facial action unit recognition Neural Information Processing Systems pp 109-17.

[21] Zhang T, Zheng W, Cui Z, Zong Y and Li Y 2019 Spatial-temporal recurrent neural network for emotion recognition IEEE Transactions on Cybernetics 49 (3) pp 839-847.