Research on Emotion Recognition Technology Driven by Deep Learning

Xiaoyan He [1, a, *], Honghu Zhang [1, b], Zhiguo Zhang [1, c], and Xiaopeng Wang [1, d]

[1] Taiyuan University of Technology, Shanxi, China;

[a, *] hexiaoyan01@tyut.edu.cn, [b] 1853698063@qq.com, [c] 2997315060@qq.com,

[d] 1240975051@qq.com

Abstract. Facial expression analysis is a crucial tool in medical diagnosis, helping doctors better understand patients' psychological states and emotional changes. Traditional methods rely on extracting and classifying facial features from images but are often affected by noise and occlusions. To enhance accuracy and efficiency, researchers have introduced convolutional neural networks (CNNs) in deep learning, which automatically extract deep features from facial images. This approach minimizes the impact of noise and occlusion, providing more precise emotion analysis to support medical diagnosis and treatment.

Keywords: Deep learning; expression recognition;feature extraction; face detection.

1. Introduction

In the era of rapid advancements in medical technology, artificial intelligence and computer vision are deeply integrated to build efficient and reliable human-computer interaction systems in healthcare. Research shows that facial expressions are crucial in doctor-patient communication, carrying over 55% of emotional and intentional information.

Deep learning has revolutionized facial expression recognition by enabling automatic feature extraction with high accuracy on large datasets. This study improves upon existing models such as VGG and residual networks. The VGG network is enhanced with a channel-spatial attention mechanism and an advanced residual module featuring squeeze-and-excitation blocks. A refinement module further boosts recognition accuracy. In the residual network, a cropping mask module augments data, the Ghost module reduces redundant parameters, and a combination of channel and multi-scale spatial attention mechanisms is used. A joint loss function optimizes inter-class margins and intra-class distances, aiding doctors in accurately interpreting patients' emotions.

2. Facial Expression Recognition

Facial expression recognition encompasses face detection, feature extraction, and expression classification. It holds significant research value across various fields. Face detection identifies the presence and location of faces, while feature extraction transforms raw data into high-level image representations and reduces dimensionality, playing a critical role in improving recognition accuracy[1]. Expression classification distinguishes expression types based on facial motion features[2].

2.1 Face Detection

Face detection is a critical step in the facial expression recognition process, with the primary goal of accurately determining the location of faces in a given sample image and obtaining the coordinates of key points[3]. Among the various algorithms used for face detection, the Boosting Algorithm stands out, as depicted in the "Boosting Algorithm Flowchart",Fig. 1. This algorithm involves an iterative process where multiple weak classifiers are combined into a strong classifier through a series of weighted votes[4]. In addition to the Boosting Algorithm, there are other main face detection algorithms that include:

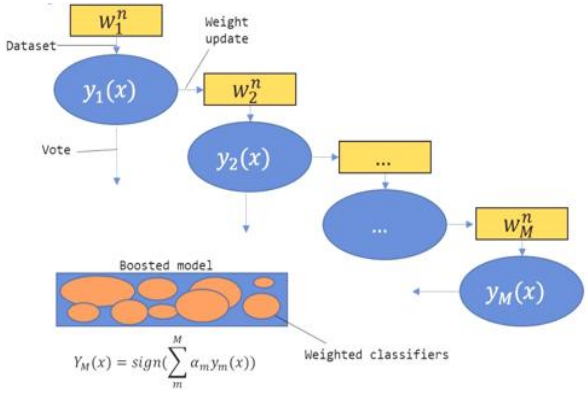

Fig. 1 Boosting Algorithm Flowchart

AdaBoost-based face detection uses ensemble learning to combine multiple weak classifiers focused on simple features (e.g., pixel intensity differences) into a strong classifier. The classifiers are iteratively trained to minimize errors, with each assigned a weight based on accuracy[5]. AdaBoost's cascaded structure allows sequential elimination of non-face regions, focusing computational resources on promising areas. This ensures high accuracy and efficiency, making it suitable for real-time applications [6].

Feature-based methods extract facial expression-related location features, offering good performance with low computational requirements. However, they need tailored feature selection and classifier training for specific scenarios, limiting generalization [7].

Deep learning methods like Cascade CNN and MTCNN use end-to-end learning to automatically extract high-level features. Cascade CNN improves detection through multiple stages, using simple models for initial screening and complex models for precise detection. These methods are ideal for large-scale datasets, while feature-based methods are better suited for resource-constrained scenarios [8],Network Structure as shown in Figure 2.

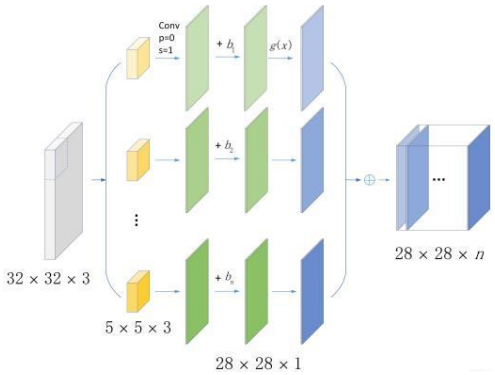

Fig. 2 Network Structure

2.2 Feature Extraction

Feature extraction is essential for accurate face detection. Preprocessing steps like grayscale conversion, histogram equalization, and normalization enhance feature representation and reduce lighting and contrast variations. Traditional methods (e.g., Gabor filters, LBP, HOG) capture local texture and shape but require manual feature engineering and may fail in complex or noisy datasets [9]. Deep learning, using CNNs, automates feature extraction, eliminating manual design and improving accuracy on large-scale datasets [10]. It also employs techniques like raw data learning, residual connections, dimensionality reduction, and hybrid approaches to optimize accuracy and robustness in face detection systems.

2.3 Expression Classification

Expression classification is the final step in the facial expression recognition process. The basic approach involves first determining the facial region in the sample image through face detection, then extracting sufficiently informative features for classification. The classification categories are divided into seven types: anger, disgust, fear, happiness, neutrality, sadness, and surprise. Based on the extracted features, the computer determines the expression category to which the input sample image belongs[11]. Currently, many researchers, both domestically and internationally, primarily use Support Vector Machines (SVM) or CNN-based classification algorithms for expression classification tasks.

3. Facial Expression Recognition Using Convolutional Network Attention Mechanisms

Currently, most deep learning-based facial expression recognition techniques aim to achieve higher accuracy by deepening the network layers or improving convolutional network models. However, simply increasing the number and width of the network layers does not significantly improve the model's ability to recognize facial expressions. In fact, it may lead to a variety of issues, such as unstable gradients, high computational complexity, feature representation bottlenecks, and network degradation, which in turn results in poorer network performance and increased error rates in expression recognition[12]. To address these challenges, researchers both domestically and internationally have proposed various network architectures, with models such as VGG, AlexNet, and ResNet commonly used as foundational networks in the field of facial expression recognition.

3.1 Network Design

The facial expression recognition algorithm uses convolutional networks with attention mechanisms and residual modules to extract features and reduce disturbances. It combines branch and main networks for precise processing and employs a shared loss function to improve accuracy.

3.2 Shallow feature extraction layer

The "shallow feature extraction layer" refers to the initial layers of a neural network that are responsible for extracting basic, low-level features from the input data, such as edges, textures, and simple patterns. These features are typically more general and not highly abstract, focusing on detecting fundamental visual elements in an image. In the context of convolutional neural networks (CNNs), shallow layers often involve simple operations like convolution and pooling, which help in capturing local structures before passing the information to deeper layers for more complex feature extraction and interpretation. These shallow layers serve as the foundation for the network to progressively learn more sophisticated and abstract features in subsequent layers.

3.3 Attention mechanism

In facial emotion recognition, conventional CNNs predict emotions based on overall facial structure but may be influenced by irrelevant components, degrading performance. The attention mechanism addresses this by reassigning feature weights through masks, focusing on critical facial regions and ignoring irrelevant information.

This chapter introduces a new attention mechanism based on CBAM: the Channel-Spatial Attention Module (CSAM). Unlike CBAM, CSAM combines channel and spatial attention mechanisms in parallel. Channel attention uses global average pooling (GAP) to compress features, while spatial attention applies convolution after both average and max pooling, preserving more channel information before fusion. This enriches the network structure and enhances key features.

3.4 Terminal feature extraction and classification layer

This paper proposes a refined model for facial detail processing. The joint loss function, as in (1), increases inter-class distance between facial expression categories, improving classification performance by optimizing feature extraction and classification simultaneously. This results in a robust model that distinguishes subtle differences between similar expressions. The model also integrates deep convolutional networks and attention mechanisms to enhance key facial feature extraction, focus on critical regions, and minimize irrelevant information. This approach achieves higher accuracy and generalization, especially for less pronounced or occluded expressions, ensuring reliable emotion recognition in diverse real-world scenarios.

$$L_{joint} = \lambda_1 L_{cls} + \lambda_2 L_{reg} + \lambda_3 L_{dist} \quad (1)$$

4. Facial expression recognition using the combination of residual networks and attention mechanisms

Different network models each have their own distinct features, but they also face some issues, such as low recognition accuracy and large model sizes. To overcome these challenges, this chapter introduces a new network model: initially, images are cropped and masked for data augmentation; then, a "compression-excitation" step is integrated into the improved Ghost model to minimize noise interference; lastly, multi-level spatial attention mechanisms are employed to enhance the model’s ability to capture finer image details. Building on this, a facial expression classification method based on images is proposed.

4.1 Network Design

This chapter proposes a new facial recognition algorithm based on facial features (Figure 3). It integrates a residual network with a segmentation mask module, Ghost module, channel attention mechanism, and multi-scale spatial attention mechanism. A random occlusion method enhances overall information utilization, while channel attention assigns weights to important features. Multi-level spatial attention expands the receptive field and extracts texture features. Additionally, a depthwise separable convolution model reduces parameters, and a compression-excitation (SE)-based method minimizes noise impact.

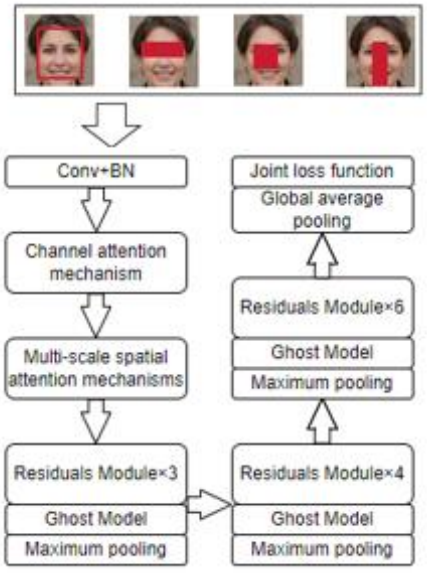

Fig. 3 Network Model Architecture

4.2 Channel Attention Mechanism

The Channel Attention Mechanism (CAM) recalibrates features by highlighting important channels and suppressing irrelevant ones. It works by first applying global average pooling (GAP) and global max pooling (GMP) to the feature map, generating two $1 \times 1 \times C$ feature maps representing average and maximum values. These maps are then compressed through a shared fully connected or convolution layer with a reduction factor r, followed by a ReLU activation function. The resulting weights from both pooling operations are merged (usually by addition) and passed through a sigmoid function to produce the final channel weights. These weights are then multiplied with the original feature map, selectively strengthening important features.

Technically, CAM enables the network to focus on the most informative channels by capturing both global context (via GAP) and distinctive features (via GMP). This dual approach prioritizes crucial features, enhancing accuracy and efficiency in tasks like facial expression recognition where subtle details are critical.

4.3 Multi-scale spatial attention mechanism

In facial emotion recognition, key features such as eyes, lips, and eyebrows are an important basis for recognition because they contain rich texture information. Subtle changes in these traits can significantly reflect the emotional state and can be easily detected by computers. However, when the sample image is cutout, some facial features are occluded, which increases the difficulty of feature point extraction. Due to the different proportions of faces in different images, the multi-scale characteristics and spatial attention mechanism can be used to effectively extract the key nodes in the images by enhancing the weight of each receptive field. The specific process is shown in Fig. 4.

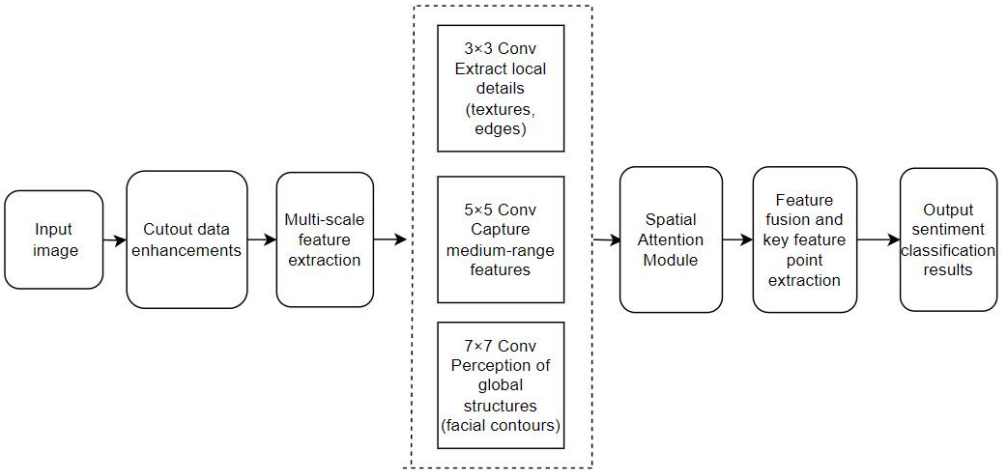

Fig. 3 Multi-scale Spatial Attention Mechanism

5. Conclusion

This paper focuses on the three core steps of facial expression recognition—face detection, feature extraction, and expression classification—comparing traditional and deep learning algorithms to provide a theoretical basis for further research. Based on this, an attention mechanism model using Convolutional Neural Networks (CNNs) is proposed. The model introduces the Channel-Spatial Attention Mechanism (CSAM) to adaptively allocate feature weights and highlight subtle differences in expression areas. It also integrates an improved Enhanced High-Order Residual Module (EHORM) to reduce parameters and enhance attention to key expression regions, improving recognition ability.

Specifically, the model uses ResNet as the backbone and incorporates Cutout Mask technology to randomly occlude image areas, enhancing robustness to occlusion and noise. The CSAM captures channel dependence and spatial information in parallel to accurately locate key expression

areas. The EHORM improves feature expression through multi-order feature fusion and reduces computational complexity. This design offers a new solution for facial expression recognition, with its effectiveness to be verified by future experiments.

References

[1] Mollahosseini, Ali, et al. "AffectNet: A database for facial expression, valence, and arousal computing in the wild." IEEE Transactions on Affective Computing 10.1 (2019): 18-31.

[2] Li, Xin, et al. "Deep facial expression recognition: A survey." IEEE Transactions on Affective Computing 10.3 (2019): 392-405.

[3] Zhang, K., et al. "Facial expression recognition via deep neural networks." Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015: 3422-3430.

[4] Zhao, Guang, et al. "Deeply learned face representations are sparse, selective, and robust." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 3297-3306.

[5] Cheng, Jiwen, et al. "EmotionNet: Fine-grained emotion cause detection with emotion-driven deep convolutional networks." IEEE Transactions on Affective Computing 7.3 (2016): 335-347.

[6] Zhou, Zhiqiang, et al. "Facial expression recognition with deep learning: A survey." Journal of Visual Communication and Image Representation 58 (2019): 37-52.

[7] Hussain, Waseem, et al. "Emotion recognition from facial expressions using deep learning." Proceedings of the International Conference on Machine Learning and Data Mining (MLDM), 2019: 110-120.

[8] Jaiswal, Abhinav, et al. "Real-time facial expression recognition using deep learning and multiscale convolutional networks." Sensors 20.1 (2020): 42.

[9] Yin, Li, et al. "Deep learning for emotion recognition on small datasets using transfer learning." Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), 2017: 73-81.

[10] Liu, Ying, et al. "Facial emotion recognition using a hybrid model of deep convolutional neural networks." Pattern Recognition Letters 103 (2018): 79-86.

[11] Xu, Zhen, et al. "Real-time facial expression recognition using convolutional neural networks." IEEE Access 7 (2019): 151971-151980.

[12] Zhou, Yuan, et al. "Facial expression recognition via a hybrid deep learning model." Journal of Visual Communication and Image Representation 56 (2018): 202-211.