Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/104/20241112

A Review of Facial Expression Recognition Based on Convolutional Neural Networks

Chenxuan Zhu

Fuzhou University Graduation, Fuzhou, Fujian Province, 350000, China

2033258676@qq.com

Abstract. The intelligent processing of human facial expressions has gained popularity due to the growth of big data and the expanding knowledge of AI and machine learning. In this paper, the author reviews the current mainstream methods for face recognition using convolutional neural network (CNN) models in deep learning, providing insights and future directions. By collecting and analyzing research findings, the characteristics, strengths and weaknesses of each model are discussed. The results indicate that CNN-based face recognition systems typically involve several steps: face detection, face alignment, face representation, face matching, and post-processing. By examining expression recognition with models like LeNet-5, AlexNet, VGGNet, and GoogleNet, the author concludes that CNNs possess features such as being data-driven, capable of feature learning, multi-level structured, adaptable to different domains, capable of real-time processing, and supporting multi-task learning. This technology has made significant strides and offers considerable potential for further development.

Keywords: CNN, facial expression recognition, machine learning, deep learning.

1. Introduction

Research on facial expression recognition began in the 1970s, with early research mainly focusing on static images and using the feature face (Eigenface) method for facial recognition. By the 1980s, researchers started exploring facial expression recognition in moving images, utilizing techniques such as the facial action coding system (FACS) and dynamic feature-based methods. The application of neural networks for facial expression identification began in the 1990s as a result of developments in computer vision and machine learning. The emergence of deep learning in the early 21st century led to important advancements in the discipline. When it came to face expression identification tests, CNN performed remarkably well. In fact, several researchers found that integrating CNNs with manually created features like LBP increased recognition accuracy. Facial expression recognition technology has advanced and become widely used in areas including intelligent monitoring, affective computing, and human-computer interaction in recent years[1]. Facial expressions play a crucial role in nonverbal communication. It can enable facial recognition systems to better understand human emotions[2]. The practical applications of facial emotion recognition are extensive and diverse[3]. This paper focuses on studies of four convolutional neural network models applied to face recognition. By collecting and analyzing research findings, the characteristics, strengths, and limitations of each model are summarized. By comparing their advantages and development purpose, the evolution for facial

DOI: 10.54254/2755-2721/104/20241112

recognition can be traced. Finally, this analysis points to future directions for the development of the face recognition model, offering valuable insights for further research.

2. Facial expression recognition

It is a technology that identifies human emotions by analyzing changes in the movement and shape of facial muscles. This technology has broad applications in fields such as human-computer interaction, game development, intelligent robotics, and security monitoring[4]. The algorithm first identifies the location of the face in an image or video stream. This can be done using pre-trained deep learning models such as MTCNN, Haar Cascade, or other facial detection algorithms. Once the face is detected, key points on the face are located to map facial features. Based on the positions and movements of these key points, potential facial expressions are analyzed. This usually involves comparing the detected facial features to predefined expression models such as FACS or EDLP. The analysis results are then used to classify facial expressions into specific emotional categories. This can be done using a classifier, like a support vector machine or neural network, trained on labeled facial expression images. Finally, the recognition results are presented in a user-friendly way, such as with text labels or emotion scores.

3. Convolutional neural networks

CNN are primarily used in tasks like image recognition, classification, and detection. A CNN typically consists of convolutional layers, pooling layers, and fully connected layers. The convolutional layer is the core component of a CNN. It involves sliding a convolutional kernel over the input image, weighting specific regions, and processing the results with an activation function like ReLU. Figure 1 illustrates how the convolution kernel extracts information from the input, and Figure 2 provides a visual representation of this process. By stacking multiple convolutional layers, CNNs can learn a variety of abstract features from the image.

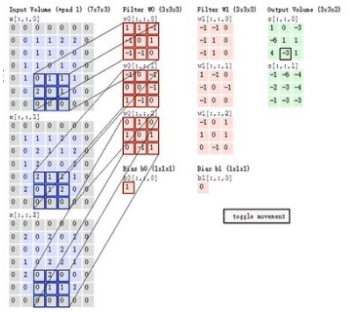

Figure 1. Convolutional layer schematic [5]

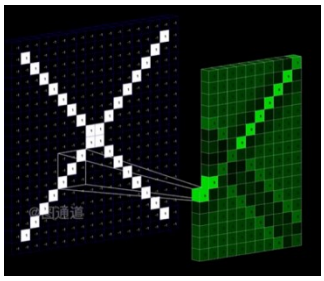

Figure 2. Intuitive diagram[6]

The pooling layer usually follows a convolutional layer, which reduces computation by reducing the resolution. Pooling operations include maximum pooling, average pooling, and mixed pooling, among others. Figure 3 shows one of its pooling methods, and Figure 4 visually illustrates the pooling layer. These operations can reduce the size of the feature map, reduce the number of parameters, expand the field of view, and maintain immutability.

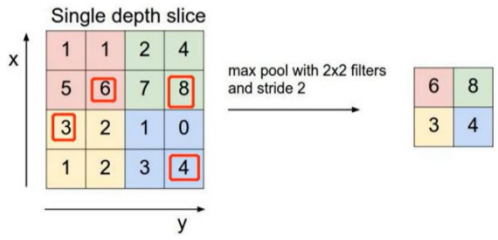

Figure 3. Maximum pooling [5]

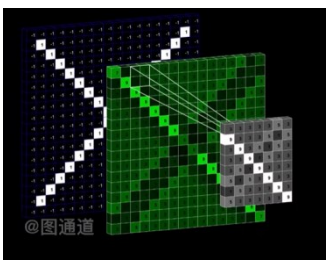

Figure 4. Intuitive diagram of pooling layer[6]

DOI: 10.54254/2755-2721/104/20241112

The fully connected layer is the last part, which merges all the feature maps into a single vector and then generates the final output via the softmax function. Prior to the fully connected layer, a flattening operation is often performed to convert the multidimensional feature map into a one-dimensional vector. Figure 5 shows a simplified model of a fully connected layer.

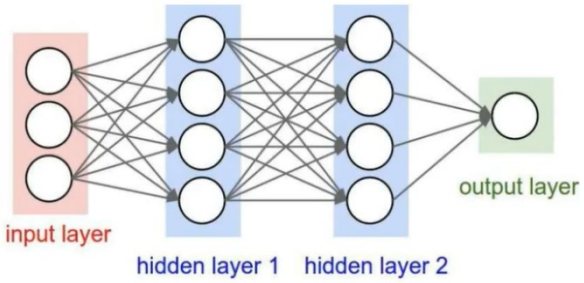

Figure 5. Simplified model of fully connected layer [5]

The dimensions of the convolutional kernel, the number of layers in the neural network, and the number of neurons in each layer must all be determined before creating a model, avoiding overfitting, and modifying hyperparameters in a convolutional neural network design.

4. Facial expression recognition based on convolutional neural network

CNN models can be divided into two categories: one is to directly use existing CNN models for fine-tuning, such as AlexNet, VGG, ResNet, etc.; the other type is CNN models specially designed for facial expression recognition, such as Deep Expression, Expression CNN, etc. These models have achieved high recognition accuracy on some publicly available facial expression datasets (such as FER2013, CK+, etc.). Figure 6 illustrates the development process, and branching of convolutional networks.

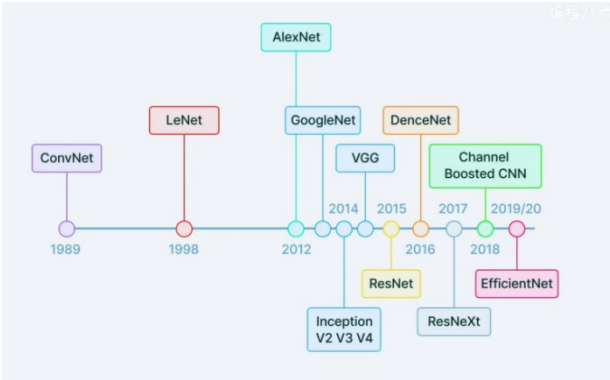

Figure 6. Convolutional neural network branching diagram [7]

4.1. Expression recognition based on LeNet-5 model

The LeNet-5-based expression recognition model consists of these parts: the input layer, the first convolutional layer, the maximum pooling layer, the second convolutional layer, the fully connected layer, and the output layer [8].

To more intuitively reflect its structure, the author introduces a simple LeNet-5-based expression recognition model:

- 1. Input Layer: Enter a face image with a size of 32x32x3.
- 2. First convolutional layer: Use 6 5x5 convolutional kernels, each with 6 output channels. Then use the ReLU activation function.
- 3. Maximum Pooling Layer: Use a 2x2 pooling window in steps of 2.

DOI: 10.54254/2755-2721/104/20241112

- 4. Second convolutional layer: Uses 16 5x5 convolutional kernels, each with 16 output channels. Then use the ReLU activation function.
- 5. Maximum Pooling Layer: Use a 2x2 pooling window with a step size of 2.
- 6. Fully Connected Layer: The output of the pooled layer is flattened into a vector and then connected to a fully connected layer containing 120 neurons. Then use the ReLU activation function.
- 7. Output Layer: The output layer contains 7 neurons (representing each of the 7 basic facial expressions: anger, disgust, fear, happiness, sadness, surprise, and neutrality) using Softmax as the activation function.

When training a model, we can use large-scale facial expression datasets (e.g. FER2013, CK+, etc.). The training process typically consists of the following steps:

- 1. Data preprocessing: cropping, scaling, and normalizing images.
- 2. Divide the training set and the test set: Typically divide the dataset into 70% training and 30% testing.
- 3. Define the loss function and optimizer: Use the cross-entropy loss function and the stochastic gradient descent (SGD) optimizer.
- 4. Train the model: update the model parameters through the backpropagation algorithm.
- 5. Test the model: Evaluate the performance of the model on the test set.

The LeNet-5 model was originally used for the recognition task of handwritten numbers and its advantages are that the structure is simple, and the construction and training costs are relatively low. It is also because of the relatively simple structure, which can be difficult to capture for complex facial expression features. Therefore, it has a relatively narrow range of applications.

4.2. Expression recognition based on AlexNet model

AlexNet is a deep convolutional neural network proposed in 2012 and won first place in the ILSVRC [9].

It is improved and optimized on the basis of LeNet, and consists of five convolutional layers, three pooling layers, and three fully connected layers.

In addition to convolution and activation operations, each convolutional layer also performs LRN local response normalization. Specifically, for each pixel on the feature map, LRN calculates the sum of squares of all pixel values of its neighbors, and then uses this sum of squares to normalize the pixels.

Through local normalization, different features in the network will have more consistent scales during the training process, thereby improving the generalization ability of the model.

The three pooling layers followed the first, second and fifth convolutional layers, respectively, and the pooling method used maximum pooling, and AlexNet used ReLU as the activation function for the first time, which helped to alleviate the problem of gradient vanishing.

To prevent overfitting, Alex Net added a dropout operation that randomly discards a subset of neurons during training, allowing the network to learn more robust features.

In addition, AlexNet also adopts data augmentation technology to increase the generalization ability of the model by randomly cropping, flipping, and color adjusting the training image.

The expression recognition based on the AlexNet model has a high accuracy rate and as long as there is sufficient training data, its accuracy is comparable to that of experienced physiognologists. However, it requires a large amount of data to train, and the cost of training is high. In addition, the AlexNet model also has some limitations, such as high computational complexity and many parameters, which is a significant challenge for developers. The model to be introduced later is an upgraded version of it designed to address its shortcomings.

4.3. Expression recognition based on VGGNet model

VGGNet is a deep convolutional neural network proposed in 2014 and has achieved excellent results in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). Since then, the model has been widely used for a variety of computer vision tasks, including expression recognition.

The main advantages of VGGNet include:

DOI: 10.54254/2755-2721/104/20241112

- 1. Simple Architecture: VGGNet adopts a simple architecture, making it easy to understand and implement.
- 2. Transfer learning capability: VGGNet can be pre-trained on large-scale datasets, and the learned features can be transferred to other visual tasks to improve performance.
- 3. Robustness: VGGNet is robust to image transformations such as scaling, panning, and rotation.

The expression recognition steps based on the VGGNet model are similar to those based on the AlexNet model, which mainly include data preprocessing, feature extraction, classification and recognition. The expression recognition based on the VGGNet model has a high accuracy rate, but it also requires a large amount of data. In addition, it also has limitations such as high computational complexity and many parameters.

4.4. Expression recognition based on GoogleNet model

The GoogleNet model is a deep learning model proposed by Google in 2014 that is mainly used for image classification tasks [10]. It can perform feature extraction from images through CNNs and then classify features using Fully Connected Layer (FC). The GoogleNet model can be used as the foundation model in the expression recognition process, and certain layers can be put on top of it to extract the expression features. For instance, topple the GoogleNet model and add a fully connected layer with as many output nodes as there are emoji types that require recognition. To determine the likelihood that an input image falls into each expression category, the fully connected layer's output can then be transformed into a probability distribution using the softmax function. It is important to note that the GoogleNet model itself is pre-trained on the ImageNet dataset, so using it directly for expression recognition may not work well. In order to improve the model performance, we can fine-tune the GoogleNet model with other expression datasets, or combine the GoogleNet model with other models to get better expression recognition.

5. Discussion

Convolutional neural network models applied to face recognition have emerged and developed rapidly over the past 20 years. The specific analysis is as follows: The LeNet-5 model was originally intended for handwriting character recognition tasks, and it is not widely used in face recognition tasks, but some of its key components, such as the convolutional layer, have laid the foundation for other advanced face recognition models. The AlexNet model has also achieved good results in face recognition tasks but it may be limited in practical applications due to its large number of parameters and high computational complexity. VGG models mainly include VGG-11, VGG-13, VGG-16, VGG-19, and other models with different depths. The VGG model has high accuracy in face recognition tasks, but the small receptive field may lead to the weak ability of the model to capture global information. The GoogleNet model introduces a new architecture called "Inception" to capture multi-scale information by using multiple convolutional kernels of different sizes in parallel. The GoogleNet model has high accuracy and efficiency in face recognition tasks, but in some cases, the generalization ability of the model may be limited.

6. Conclusion

In conclusion, CNNs can effectively capture feature representations from facial images, leading to significant performance improvement across various face recognition tasks. Typically, a CNN-based face recognition system involves several steps: face detection, face alignment, etc. By examining models like LeNet-5, AlexNet, VGGNet, and GoogleNet, it’s clear that face recognition models are advancing rapidly, becoming more precise, detailed, and structured. However, despite the substantial progress, challenges remain, such as handling complex variations like changes in pose and lighting, as well as managing the computational complexity when working with large-scale datasets. To address these challenges, researchers have introduced improvements like more intricate network structures, the use of harder negative samples during training, and online hard sample mining techniques. Future research will focus on tackling face recognition in dynamic environments, designing more efficient

DOI: 10.54254/2755-2721/104/20241112

CNN models and algorithms, and exploring loss functions and training strategies that are better suited for face recognition tasks.

References

[1] H. Ali et al. Facial emotion recognition using empirical mode decomposition Expert Systems with Applications (2015).

[2] Pranav, E.; Kamal, S.; Chandran, C.S.; Supriya, M. Facial emotion recognition using deep convolutional neural network. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 March 2020; pp. 317–320.

[3] R. Jack et al. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time Curr.Biol. (2014).

[4] Li, Yong, et al. "Occlusion aware facial expression recognition using CNN with attention mechanism." IEEE Transactions on Image Processing 28.5 (2018): 2439-2450.

[5] MYVision_mayishijue. Deep Learning Overview[EB/OL]. (2023.1.9)[2024.8.24].bilibili.com

[6] Akhand M. A. H., et al. "Facial emotion recognition using transfer learning in the deep CNN." Electronics 10.9 (2021): 1036.

[7] Zadeh, Milad Mohammad Taghi, Maryam Imani, and Babak Majidi. "Fast facial emotion recognition using convolutional neural networks and Gabor filters." 2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI). IEEE, 2019.

[8] Zhang, J., Yu, X., Lei, X., & Wu, C. A novel deep LeNet-5 convolutional neural network model for image recognition. Computer Science and Information Systems, 19(3), 1463-1480. (2022).

[9] Yuan, Z. W., & Zhang, J. Feature extraction and image retrieval based on AlexNet. In Eighth International Conference on Digital Image Processing (ICDIP 2016) (Vol. 10033, pp. 65-69). SPIE. (2016, August).

[10] Anand, R., Shanthi, T., Nithish, M. S., & Lakshman, S. Face recognition and classification using GoogleNET architecture. In Soft Computing for Problem Solving: SocProS 2018, Volume 1 (pp. 261-269). Springer Singapore. (2020).