Literature Review of Face Expression Recognition

1. Introduction

Facial expression recognition(FER) is a technology aimed at inferring emotional states15or expressions by analyzing and interpreting human facial expressions. This technology employs computer vision and pattern recognition techniques to determine the emotions being experienced by an individual. It accomplishes this task by identifying and analyzing variations in facial features, including but not limited to happiness, sadness, anger, surprise, etc. FER was first proposed by Paul Ekman (1972)[2], who elaborated a systematic classification and description of facial expressions, and proposed seven basic facial expressions: anger, disgust, fear, happiness, sadness, surprise, and contempt. Recently, the importance of facial emotion recognition (FER) has significantly increased due to its wide-ranging applications in diverse fields. These include human-computer interaction, healthcare, social robotics, and security systems[7]. FER poses a challenging problem as it involves the recognition and interpretation of human emotional states conveyed through facial expressions. The ability to accurately analyse and understand facial emotions holds immense potential for improving communication, enhancing user experiences, enabling personalized healthcare, and developing more intuitive human-machine interfaces. As a result, FER research and development continue to advance to unlock its full potential in diverse applications.

With the development of computer vision technology, expression recognition has attracted more and more attention. Initially, researchers employed Principal Component Analysis (PCA) to extract main features from facial image datasets and represent them as feature vectors.[6][9][10] However, these methods had limitations in handling variations in lighting, pose, and expression, leading to lower robustness. To overcome these limitations, researchers turned to machine learning methods for facial emotion recognition. Support Vector Machines (SVM) and Artificial Neural Networks (ANN) were explored as alternative approaches. [3][5][8] SVM-based real-time recognition, automatic feature extraction, and classification using ANN, as well as multi-classification using discrete wavelet transform and SVM, were proposed in different studies. These machine learning methods exhibited better adaptability compared to feature extraction approaches. However, challenges such as high computational complexity, sensitivity to noise, and parameter selection sensitivity persisted. In recent years, the rapid development of deep learning technology has brought new breakthroughs in the field of FER.[1][4][13][14] Various researchers have suggested differing approaches and models for facial emotion recognition. These include the hybrid deep learning CNN-RNN model [4] , regional attention network (RAN)[13], transfer learning-based facial emotion recognition system[1], and three-dimensional convolutional neural network and convolutional long-term short-term memory

(LSTM)[11]. Hybrid models, for FER and emotion recognition, have achieved remarkable progress in feature learning, spatio-temporal modeling, region attention, and transfer learning. Compared with machine learning methods, deep learning can automatically acquire features and adopt some special structural layers to resist the influence of noise, which improves robustness and accuracy.

Although the existing technology has been quite developed in the field of FER, there are still some unsolved problems. These include the challenges of small samples and unbalanced data, and the challenges of accurate recognition of diverse and dynamic facial expressions. Therefore, further research will focus on filling these knowledge gaps to develop more effective solutions.

The structure of this paper is shown in Figure.1. This paper is structured as follows: Firstly, it introduces the mainstream databases used for expression recognition. Then, based on the classification standard of expression recognition method, it provides an overview of FER research in terms of feature extraction, machine learning methods, and recent advancements. Next, aiming at the gaps in these studies, ideas and suggestions for future research are provided.

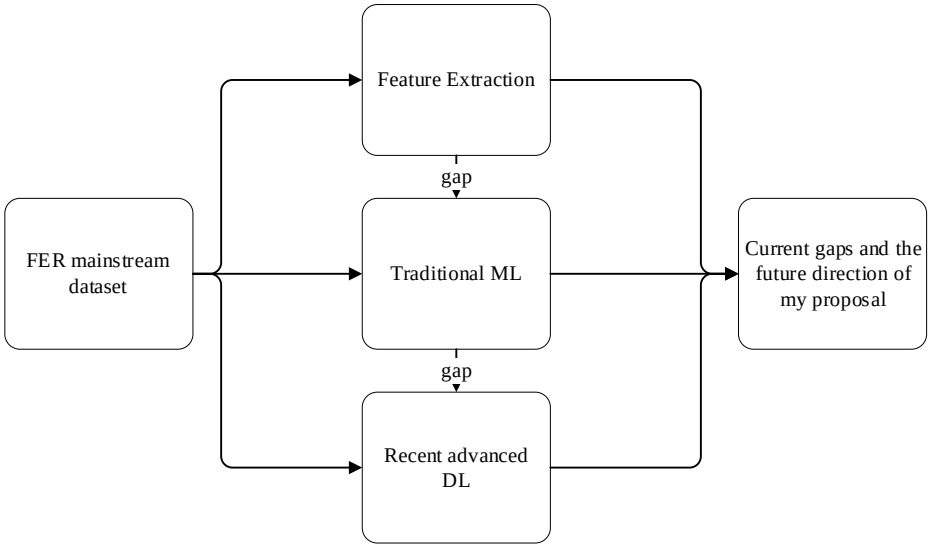

Figure 1 Article Structure Arrangement

2. Mainstream FER Database

A high-quality dataset is a crucial component in FER, which plays a key role in ensuring the validity and accuracy of the model. The datasets offer ample training and test samples, creating a foundation for model learning and evaluation. With a diverse range of data at our disposal, various emotional and facial expressions can be fully encompassed, enabling models to better comprehend and identify human emotions. In addition, a high-quality dataset can reduce bias and imbalance problems, ensuring the robustness of the model in various populations and scenarios. Table1 below summarizes some mainstream facial emotion recognition datasets,

including the size of the dataset, the image categories contained in the dataset, and the description of the dataset. In order to select the appropriate training data set for subsequent research.[7][12]

Dataset	size	type	description
CK+	593	Image	This process captures transitions from neutral to peak expressions. Evaluation involves selecting peak expression frames and the initial frame. Subjects are divided into n groups for person-independent n-fold cross-validation, typically using values like 5, 8, or 10.
MMI	740/2,900	Image/Video	Contains 740 images and 2,900 videos, onset-apex-offset labeled sequences.
JAFFE	213	Image	Images of 10 Japanese females, with each individual having 3-4 images for six fundamental facial expressions and one image for a neutral expression.
FER-2013	35,887	Image	These images are mainly from pictures on the Internet and screenshots of YouTube videos
AFEW 7.0	1,809	Video	Includes video clips gathered from diverse films that feature natural expressions, different head orientations, obstructions, and fluctuations in lighting conditions.
SFEW 2.0	1,766	Image	Created by selecting static frames from the AFEW (Acted Facial Expressions in the Wild) dataset.
Multi-PIE	755,370	Image	From 337 subjects, taken from up to four recording sessions and captured under 15 different viewpoints and 19 illumination conditions.
BU-3DFE	2,500	Image	Captured from 100 individuals. Each subject is prompted to display six facial expressions with varying intensities.
BU-4DFE	606	Image	3D facial expression sequences, comprising approximately 60,600 frame models and support for studying expressions across time.
Oulu-CASIA	2,880	Image	Image sequences obtained from 80 subjects
KDEF	4,900	Image	70 Swedish participants.
EmotioNet	Over 1,000,000	Image	From the internet, 950,000 annotated automatically, and 25,000 manually annotated with 11 AUs.
RAF-DB	29,672	Image	A real-world dataset highly diverse facial images from the Internet.
AffectNet	Over 1,000,000	Image	Obtained by searching the Internet with emotion-related tags
ExpW	91,793	Image	A dataset of 91,793 facial images downloaded using Google Image Search
4DFAB	Over 1,800,000	Image	A total of 180 subjects' high-resolution 3D faces were captured in four separate sessions over a period of five years.
AT&T	400	Image	Contains face images of 40 persons, with 10 images of each.

Oulu Physics	125	Image	Includes frontal color images of different faces
XM2VTS	100GB	Video	Consists of 1000 GBytes of video sequences
Yale	165	Image	Consists of 15 individuals, and each person has 11 grayscale frontal face images.
Yale B	5760	Image	Total of 5760 images

Table 1 Mainstream databases

3. Feature Extraction-based methods

The early methods of FER mainly used feature extraction. Generally, the steps of this method are as follows: First, a training data set is collected, in which each sample contains a face image and the corresponding label (expression category). Next, convert the image into a vector representation, for example by flattening the pixel values into a one-dimensional vector. Then, PCA algorithm is applied to reduce the dimensionality of these vectors. The core idea of PCA is to find the principal components that contain the most information in the data, and express the data by retaining the features with the largest variance. Therefore, in expression recognition, the first few principal components (the ones with the largest eigenvalues) are selected as the features of expression recognition. Below I will introduce several past related studies and analyze their strengths and weaknesses.

3.1 FER Based on PCA and Karhunen

Sirovich et al (1987) proposed a face recognition method based on PCA[10]. The purpose is to explore a low-dimensional vector representation method for characterizing, recognizing and distinguishing individual facial features in pictures. The authors discovered a low-dimensional vector representation based on Karhunen-Loeve expansion and principal component analysis called eigenpictures. The eigenpictures are determined by averaging the covariance of the collection of pictures and determining the eigenvectors of the corresponding symmetric matrix. Using a set of eigenpictures, a face of person can be represented by a small set of numbers, significantly reducing the amount of data required for classification.

Although the paper is older, it has a groundbreaking character. It proposes for the first time the use of Karhunen-Loeve's extended pattern recognition and principal component analysis techniques to reduce the complex face feature data to a dimensionality level, thereby significantly increasing the calculation speed of face analysis. The appearance of this method laid the foundation for the follow-up research. However, this paper also has some limitations due to the immature technology at that time. For example, it cannot handle large-scale face data, and has high requirements for lighting and pose. This limits the applicability of this method in practical applications. In general, the contribution of this paper is to create a novel method, which brings new ideas and technical means to the field of face analysis by introducing Karhunen-Loeve extension and principal component analysis. Although it has some limitations, it provides an important foundation for subsequent research and has had a profound impact in the fields of face recognition and pattern recognition.

The study by Kirby et al.(1990)[6] provided an overview of techniques and methods in pattern analysis and machine intelligence for recognizing and localizing two-dimensional

objects. It covered topics such as autoregressive modeling, shape matching, and polynomial fitting. The paper investigated how natural symmetry in patterns could enhance these methods. Natural symmetry referred to symmetric properties within a pattern set that could be used to improve recognition and localization. Operations like rotation and flipping could achieve these symmetry properties. For example, in face recognition, exploiting the left-right and top-bottom symmetry of faces could enhance accuracy.

The conclusion of the paper was that utilizing natural symmetry could improve the accuracy of PCA in the field of face recognition. The authors conducted experiments and found that by introducing rotated and flipped face images into the training set, the recognition rate of PCA on the test set significantly increased. This indicated that leveraging natural symmetry could enhance the diversity of the training set and improve the generalization ability of model.

This paper is ground breaking in that it introduces the use of natural symmetry in processing the training set, which is an enhancement compared to previous methods that solely relied on PCA for feature analysis. By leveraging natural symmetry, the paper demonstrates a significant improvement in recognition rates on the test set. This approach not only enhances the diversity of the training data but also takes advantage of the inherent patterns and structures present in the data. As a result, it leads to a more robust and accurate recognition system. Future research will also use some data enhancement techniques to improve the diversity of data. However, the study still has previous limitations, such as being sensitive to the pose and lighting conditions of the face when processing facial expressions. If the facial pose or illumination changes, it may lead to inaccurate feature extraction and affect the accuracy and robustness of expression recognition.

3.2 FER Based on View-Based and Modular Eigenspaces

Pentland et al. (1994)[9] implemented the study, which aimed to explore the use of eigenfaces for large-scale face recognition and interactive search. It proposed a technique based on perspective and modular feature space to improve the accuracy and robustness of face recognition. Additionally, the paper aimed to investigate the performance of face recognition under different viewpoints.

The paper found that decomposing face images into multiple modules and utilizing eigenfaces to describe each module could enhance the accuracy and robustness of face recognition. Furthermore, the method took into account the challenge of face recognition under different viewpoints and introduced a technique based on viewpoint multi-observer feature space to tackle it. The experimental results highlighted the effectiveness of the proposed method in large-scale face recognition and interactive search.

I totally agree with the approach of decomposing the face into multiple modules and extracting features from different perspectives. This method allows for obtaining a greater variety of features, which improves the accuracy and robustness of face recognition. By decomposing the face and extracting features separately, we can capture more diverse face information, including details and features from various viewpoints. Compared to previous studies, this multi-module feature extraction method enables the face recognition system to better handle changes in lighting, pose, and expression, leading to enhanced reliability and stability in recognition. Additionally, extracting features from different perspectives enhances the ability of system to recognize faces from multiple angles. Of course, this method still has

some limitations. Although the paper proposes a multi-viewer feature space method to deal with face recognition problems under different viewing angles, there may still be some extreme viewing angles, such as extreme side faces or overlooking angles, which may still be challenging for this method sex.

3.3 Discussion and Gap

The above researches show the research based on feature extraction in the field of face recognition, which improves accuracy and robustness by decomposing and describing facial features. They introduce different techniques and methods, such as Karhunen-Loeve extension, principal component analysis, natural symmetry, etc., to deal with the problem of feature extraction and recognition of face data. By extracting feature information from different viewing angles and exploiting natural symmetry, these methods have achieved remarkable results, improving the performance and stability of face recognition systems. At the same time, the introduction of these methods provides an important basis for subsequent research, and has had a profound impact on the fields of face recognition and pattern recognition. However, these methods still have the following problems.

1. Data requirements: Feature extraction-based methods usually require a large amount of labeled data to build accurate feature models. This can lead to very slow training.
2. Feature representation: Feature extraction methods may not be able to adequately represent facial emotion information. Using PCA in the feature extraction process may lose some important details or fail to capture subtle changes in emotional expression.
3. Model interpretability: The method based on feature extraction may be relatively black box, and it is difficult to explain how the model arrives at the result.
4. Algorithmic complexity: Feature extraction-based methods may involve complex image processing and feature extraction algorithms, requiring higher computing resources and time.

4. Machine learning-based methods

Machine learning refers to the use of statistical and mathematical methods for training and building machine learning models. Feature engineering and model selection are relied upon, typically involving classic supervised and unsupervised learning algorithms. Some of the more commonly used machine learning methods are listed in Table 2. The training process of machine learning algorithms is usually based on labeled training data, aiming to learn patterns and regularities from it in order to predict or classify new unlabeled data. In the field of FER, many researchers have conducted a series of related research based on machine learning. Below I introduce three related studies and describe their strengths and weaknesses.

supervised	linear regression	logistic regression	decision trees	support vector machines
unsupervised	clustering	dimensionality reduction	association rule mining	anomaly detection

Table 2 commonly used ML method

4.1 FER Using Artificial Neural Networks

Gargesha et al. (2002)[3] conducted research on using artificial neural networks for FER. They analysed seven basic human expression types: neutral, happy, sad, disgusted, angry, surprised, and fearful. The authors employed a multilayer perceptron (MLP) and a radial basis function network (RBFN) to classify these expressions. They used automatic feature extraction for most features, but manually extracted important feature points such as eyes, eyebrows, and mouth. These extracted features, along with the geometric data derived from contour points of the expression and corresponding neutral pictures, were input into the artificial neural network (ANN) for classification. Ultimately, the authors achieved a classification accuracy of 73% using the MLP network on the JAFFE dataset.

The advantage of this paper is that it applies ANN to expression recognition, which allows for automatic feature learning and extraction. This approach is more comprehensive than manual feature extraction and captures important feature information for facial expressions. Of course, this paper also has some limitations. For example, the automatic feature extraction method only considers the positions of eyes, eyebrows and mouth, and does not cover other facial feature points. This may cause the system to miss other important facial features when recognizing expressions.

4.2 FER Using Support Vector Machines

Michel et al. (2003) [8]conducted a study aiming to introduce an emotion recognition method based on FER in real-time video. The authors performed face localization and feature extraction using an automatic facial feature tracker. Specifically, the tracker extracted the locations of 22 facial features from a video stream and computed the displacement of each feature between a neutral frame and a representative frame of expression. Figure 2 below shows an example of a video stream. These displacements were then used as input to train an SVM classifier to recognize previously unseen expressions. Experimental results showed that the method performed well in real-time emotion recognition.

Figure 2 Locating and Tracking Facial Features Over a Series of Video Frames

In my opinion, the strength of this research is that using the facial recognition tracker method, it solves the impact of gesture on expression recognition, which is a very impressive method. This method can avoid to a certain extent the problem that facial expressions cannot be accurately recognized due to improper user posture. For my follow-up research, I can also borrow this method and use the facial recognition tracker to collect a more accurate FER dataset. Furthermore, the paper proposes an innovative approach by computing the displacement of each feature between the neutral and expression peak frames, and then trains these displacements together with the expression labels as the input of a Support Vector Machine (SVM) classifier and categories. Compared with the traditional method of directly extracting features, this method is more innovative. By considering the changes of features over time, the dynamic characteristics of expressions can be captured, which improves the accuracy of emotion classification. This idea plays an important role in promoting the research in the field

of emotion recognition, and provides a new idea for developing more effective emotion recognition algorithms. The limitation of the research is that there are still some errors and difficulties in the real-time recognition of facial expression emotions, such as head movement and insufficient camera capture may affect the classification accuracy. In addition, due to the complexity and variability of facial expressions, the method may require more training data and more sophisticated feature extraction and tracking techniques to achieve higher accuracy.

An automatic FER method based on wavelet transform and support vector machine is introduced in this research (Kazmi & Jaffar, 2012)[5]. The features of facial images are extracted using discrete wavelet transform in this method, and these features are used for training seven parallel support vector machines. Each SVM is trained to recognize a specific facial expression and is most sensitive to that expression. Multi-classification is achieved by employing a one-vs-many approach for binary classification. The outputs of all SVMs are combined using the max function. The principle is shown in Figure 3. The method was tested for multiple classification efficiency on static images from the public Japanese female facial expression database, and the results obtained good results with average accuracy of FER ranging from 81.67% to 96.00%.

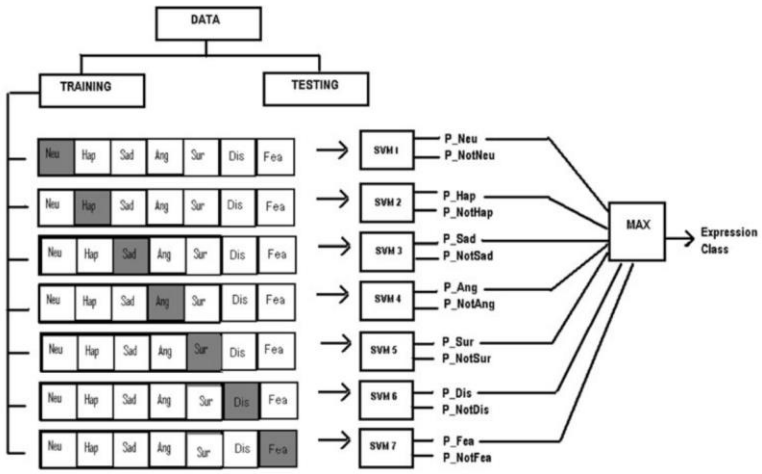

Figure 3 classifier principle

This paper has two advantages. Firstly, it uses wavelet decomposition for feature extraction, which helps in detecting subtle changes around the mouth and eyes by extracting information of different frequencies and scales. Secondly, it employs seven vector machine classifiers, each trained to recognize a specific expression. The combination of probabilities output by these classifiers, according to the maximal rule, leads to more accurate results. I agree with this research approach, and I might consider combining more advanced methods with wavelet decomposition feature extraction in future research. In this study, a potential limitation is that only the JAFFE database was used, which mainly contains expression data of Japanese women. This may lead to limited expression recognition performance of the algorithm in other racial, gender and cultural backgrounds. In order to improve the generalization ability and practicability of the algorithm, future research can consider using a wider and more diverse dataset to cover the expression characteristics of different groups of people.

4.3 Discussion and Gap

Artificial neural networks and facial feature trackers are used in machine learning-based studies to extract expressions from real-time videos, while wavelet transforms and support vector machines are employed for FER. Compared with previous feature extraction-based methods, machine learning-based methods allow for the use of a variety of features to represent facial expressions, not limited to eigenfaces. Information about color, texture, shape, etc., can be included in these features, enabling a more comprehensive description of subtle changes in facial expressions. Furthermore, high-dimensional data can be processed by machine learning-based methods, and model effects and computational efficiency can be improved through appropriate feature selection and dimension reduction techniques. These complex data can be better modeled and classified by machine learning-based methods than by eigenfaces. These studies provide a variety of solutions for the field of FER through different methods and technologies, and provide ideas and foundations for the subsequent use of more advanced methods to deal with expression recognition. Of course, there are many gaps in machine learning-based methods,

1. Reliance on feature extraction: Machine learning-based relies on manual feature engineering, which can be time-consuming, subjective, and limiting.
2. Challenges with large-scale data: Traditional methods struggle with large datasets, as they often require loading all data into memory for training.
3. Model complexity and parameter adjustment: Manual selection of models and tuning parameters can be challenging, especially for non-experts.
4. Data sparsity and generalization: Facial expression data can be high-dimensional and sparse, posing difficulties for machine learning-based methods in processing and generalizing from such data.

5. Recent Advances Deep Learning Methods

Deep learning has revolutionized the field of FER, enabling significant advancements in accurately interpreting and analyzing human emotions. By leveraging deep neural networks, these advanced models are capable of automatically learning and extracting intricate features from facial images, allowing for more precise emotion detection. Deep learning models for FER, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated remarkable performance in capturing subtle nuances and variations in facial expressions. These models can effectively learn hierarchical representations of facial features, capturing both global facial structures and local details. This enables them to detect and classify a wide range of emotions, including happiness, sadness, anger, surprise, and more. Below I will introduce four recent and relatively advanced related studies.

5.1 FER using Convolutional Neural Network

Wang et al. (2020)[13] conducted a study to address the challenges of automatic FER, especially in realistic scenes with occlusions and pose changes. To overcome these difficulties, they proposed Region Attention Network (RAN) and Region Bias Loss (RB-Loss) function. RAN aggregates and embeds various regional features produced by a backbone convolutional neural network into a compact fixed-length representation, and learns attention weights for each region through end-to-end learning. Additionally, it is then used to encourage attention to the

most important areas.

The schematic diagram is as follows,

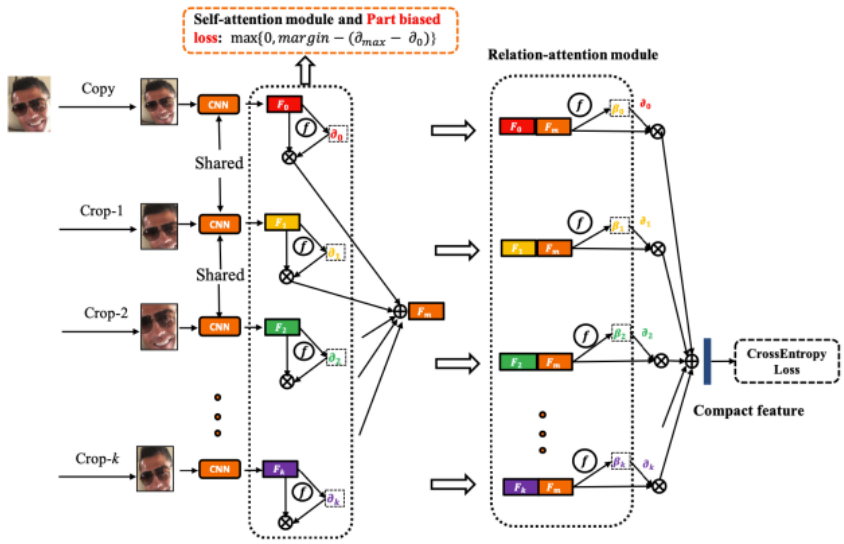

Figure 4 schematic diagram of RAN

The results of this study demonstrate the robustness of the proposed RAN and RB-Loss functions in overcoming challenges associated with pose and occlusion variations in real-world FER scenarios. In my opinion, the method not only performs well on benchmark datasets, but also has great potential for practical applications.

In my opinion, the method not only performs well on benchmark datasets, but also has great potential for practical applications. However, the limitation of this study lies in the potential issues associated with the training datasets used, namely FERPlus, AffectNet, RAF-DB, and SFEW. These datasets may have limitations such as insufficient data size and the presence of erroneous labels (noise).

My project will build on the research and methodologies in the field of FER, particularly drawing inspiration from the work of Wang et al. (2020) on the Regional Attention Network (RAN) and Region Bias Loss (RB-Loss) methods. These studies serve as a foundation and guidance for selecting appropriate methodologies and technical approaches. Furthermore, taking into account the limitations identified in the aforementioned datasets, my research aims to extend the existing knowledge by exploring more reliable and larger-scale datasets.

5.2 FER Using Transfer Learning in the Deep CNN

According to Akhand et al. (2021)[1], the purpose of their research is to propose a facial emotion recognition system based on deep convolutional neural network (CNN) and transfer learning, which can effectively identify the emotional states corresponding to different facial expressions, and through the pre-training model Applying transfer learning can reduce the required development effort. After experiments, the authors concluded that CNN-based FER systems outperform traditional geometry- and appearance-based methods in terms of accuracy and speed. These findings underscore the effectiveness of CNNs in the field of FER.

The strength of this study is that it proposes a FER system that combines deep CNNs and transfer learning to reduce development effort. There are also many limitations of this study,

for example, the ability to process images from uncontrolled environments or video sequences needs further optimization. Furthermore, while the proposed method exhibits superior performance on benchmark datasets with frontal and side views, there is still room to optimize the parameters of specific DCNN models for each dataset to potentially improve system performance.

5.3 FER Using Mix-model

Jain et al. (2018)[4] implement study titled "Hybrid deep neural networks for face emotion recognition" with the aim of investigating the application of deep learning techniques in recognizing emotions from facial expressions. Specifically, the research aimed to address the question of whether a hybrid model combining Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) can accurately recognize facial expressions and predict emotions. The findings of the study concluded that the implementation of a deep learning model, particularly the hybrid CNN-RNN model, yielded favorable results in emotion recognition through facial expressions. By leveraging the capabilities of CNN and RNN, this model demonstrated the ability to automatically learn relevant features and outperformed previous methods utilized in emotion recognition (ER). Rigorous testing and evaluation were performed under diverse circumstances and with various hyperparameters, providing evidence that the fusion of CNN and RNN significantly enhanced the overall accuracy of emotion detection.

The advancement of this paper is that it innovatively proposes a hybrid CNN-RNN model, which improves the accuracy of FER. However, this study did not systematically compare and select different model structures, hyperparameters, or optimization algorithms that may affect the performance of emotion recognition. Therefore, future research can conduct horizontal comparative experiments to tune models with different hyperparameters. This allows for comprehensive evaluation of different parameter settings and helps researchers choose the optimal model configuration.

The study by Singh et al. (2023)[11] aims to use a novel FER pipeline to recognize facial expressions in videos using a fusion of 3D Convolutional Neural Networks (3D-CNN) and Convolutional Long Short-Term Memory (ConvLSTM). Figure 5 briefly introduces the structure of the model. The authors compare their results with existing FER techniques that utilize static facial images and speech modulations to recognize emotional states, and provide examples of FER in various applications such as fatigue detection, smart mirrors, and face recognition systems.

Figure 5 hybrid 3D-CNN & ConvLSTM model

Experiments have proved that compared with the existing FER technology that uses static facial images and voice modulation to identify emotional states, the neural network model has fewer parameters, and it has three facial expression datasets of CK+, SAVEE and AFEW. A higher accuracy is achieved.

The strength of this paper is that it proposes a novel method combining 3D-CNN and

ConvLSTM, which can better process spatial and temporal information in videos. Compared with existing deep learning models, the hybrid model has fewer parameters, but can achieve comparable accuracy. Therefore, in the follow-up research, we can learn from the design ideas of this mixed model. However, due to the difficulty in obtaining facial expression data and the complexity of labeling, this study may face the problem of insufficient data. For deep learning models, large-scale and diverse data is critical to improve performance. However, in this study, there is no mention of how to deal with insufficient data. Therefore, in future research, methods to solve the problem of insufficient data need to be explored to further improve the performance and reliability of FER models.

5.4 Discussion and gap

The above studies propose innovative deep learning methods and techniques in the field of FER. Compared with traditional machine learning-based methods, deep learning methods have the following advantages: First, deep learning models can automatically learn and extract features from raw data through end-to-end learning, reducing the need for manual feature engineering. Secondly, deep learning methods can use technologies such as optimized computing graphs and distributed computing to efficiently process large-scale data sets and improve training efficiency. In addition, deep learning models can handle more complex patterns and features through hierarchical structures and automatic learning capabilities, reducing the work of manual design and adjustment of model structures and parameters. Therefore, these deep learning methods have achieved high accuracy and robustness in FER, providing an important basis and guidance for further research and application in this field.

However, there are still some gaps in the field to be resolved,

1. Insufficient data: Due to the difficulty in obtaining facial expression data and the complexity of labeling, deep learning models may face the problem of insufficient data. Large-scale and diverse data are critical to improving performance.

2. Image occlusion and light issues: occlusion may cause key expression features to be uncaptured, and changes in light will affect image quality.

3. Parameter selection and optimization: Some of the methods mentioned in the research did not systematically compare and select different model structures, hyperparameters or optimization algorithms.

6. Future work

Based on the above, the following suggestions are made for future work:

In order to solve the problem of insufficient data, data augmentation technology can be considered to expand the training data by synthesizing images or introducing other data sources. Additionally, collecting as large and diverse a dataset as possible is critical to improving performance. In order to deal with the impact of image occlusion and illumination changes, image inpainting or restoration methods can be used to deal with occluded parts, and meanwhile combine information from other sensors or modalities to assist expression recognition. This can improve the robustness and generalization ability of the model, making it suitable for practical application needs. For the choice of model structure, hyperparameters and optimization algorithms, systematic comparison and evaluation can be carried out. Comprehensively

evaluating the effects of different parameter settings through horizontal comparison experiments and tuning different hyperparameters helps researchers choose the best model configuration.

In addition, research in the following areas can also be considered: improvement of model interpretability, application of transfer learning and domain adaptation, fusion of multimodal information, exploration of online learning and incremental learning, etc. These directions will further promote the research and application development in the field of FER, improving its accuracy, robustness and practicality.

7. Conclusion

This review mainly examines the relevant literature in the field of FER. First, we reviewed commonly used facial expression datasets to select suitable databases for subsequent studies. This paper then analyzed different FER methods and algorithms, including those based on eigenfaces, machine learning, and deep learning. This paper found that deep learning methods have achieved remarkable results in FER tasks, and are currently the most advanced and mainstream FER methods. However, although the field of FER has developed very vigorously, there are still some challenges and problems, including the uneven quality and insufficient data of the expression database, and the influence of factors such as illumination changes, expression diversity, and pose changes on the recognition accuracy. influence, and how to choose appropriate model parameter selection and optimization parameters.

The significance of this study lies in the in-depth understanding of the development and application status of FER, which provides important reference and guidance for further research and application. In view of the current problems, future work in FER should focus on addressing the problems of insufficient data through data augmentation, handling image occlusion and illumination changes by employing image inpainting and fusion with other modalities, and conducting systematic comparisons to optimize model structures, hyperparameters. In short, FER has important application value in the fields of human-computer interaction and affective computing. By solving the current problems and challenges, expanding research directions, and furthering the understanding and application of FER technology, the effectiveness and benefits of FER can be improved in practical applications.

Finally, here is a statement that I used chatgpt to polish and modify sentences in this paper.

Appendix

[1] Akhand, M. A. H., Roy, S., Siddique, N., Kamal, M. A. S., & Shimamura, T. (2021). Facial emotion recognition using transfer learning in the deep CNN. Electronics, 10(9), 1036.

[2] Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). Emotion in the human face: Guidelines for research and an integration of findings. Pergamon Press.

[3] Gargesha, M., Kuchi, P., & Torkkola, I. D. K. (2002). Facial expression recognition using artificial neural networks. Artif. Neural Comput. Syst, 8(4), 1-6.

[4] Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., & Zareapoor, M. (2018). Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters*, *115*, 101-106.

[5] Kazmi, S. B., & Arfan Jaffar, M. (2012). Wavelets-based facial expression recognition using a bank of support vector machines. Soft Computing, 16, 369-379.

[6] Kirby, M., & Sirovich, L. (1990). Application of the Karhunen-Loeve procedure for the characterization of human faces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12(1), 103-108.

[7] Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. IEEE transactions on affective computing, 13(3), 1195-1215.

[8] Michel, P., & El Kaliouby, R. (2003, November). Real-time facial expression recognition in video using support vector machines. In Proceedings of the 5th International Conference on Multimodal Interfaces (pp. 258-264).

[9] Pentland, A., Moghaddam, B., & Starner, T. (1994). View-based and modular eigenspaces for face recognition.

[10] Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. Josa a, 4(3), 519-524.

[11] Singh, R., Saurav, S., Kumar, T., Saini, R., Vohra, A., & Singh, S. (2023). Facial expression recognition in videos using hybrid CNN & ConvLSTM. International Journal of Information Technology, 1-12.

[12] Tolba, A. S., El-Baz, A. H., & El-Harby, A. A. (2006). Face recognition: A literature review. International Journal of Signal Processing, 2(2), 88-103.

[13] Wang, K., Peng, X., Yang, J., Meng, D., & Qiao, Y. (2020). Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 29, 4057-4069.

[14] Yang, D., Alsadoon, A., Prasad, P. C., Singh, A. K., & Elchouemi, A. (2018). An emotion recognition model based on facial recognition in virtual learning environment. Procedia Computer Science, 125, 2-10.