Dean&Francis

ISSN 2959-6157

A Comprehensive Study of Facial Expression Recognition

Peilun Li

Marianapolis Preparatory School, Thompson, CT, USA
peilunli31@gmail.com

Abstract:

Facial expression recognition (FER) is a key area of research interest in the field of human-computer interaction. It can make an intelligent system feel and respond to humans' emotions. First, this review introduces FER's important significance in diverse domains such as healthcare, smart environments, and human-robot interaction, where emotion-aware systems can enhance communication efficiency. The review analyzes key technical components of FER systems, including dataset acquisition and preprocessing, feature extraction mechanisms, and classification models. It also discusses transfer learning, dropout regularization, and optimization strategies to improve model performance and reduce overfitting. In addition, it evaluates common performance metrics in FER research and points out the model's limitations and advantages in different emotions, such as anger, happiness, sadness, surprise, fear, and hate. It proposes data augmentation and advanced network designs to solve problems in FER, including variations in lighting, pose, and cultural divide. Finally, summarizes and points out research interests in the future, such as the fusion of multimodal data and the development of lightweight models for real-time applications. This review can promote the use of FER in daily life. Provides reference to researchers and workers in FER, and helps to explore innovation to push the development of emotional identity tech.

Keywords: FER, Deep Learning, Traditional Machine Learning, Model Interpretability

1 Introduction

In the field of healthcare, FER tech can become a doctor's right-hand man. For example, for people with mental illnesses such as depression and autism, their emotional expressions are normally abnormal. Traditional diagnosis methods depend on patient self-report and the doctor's subjective judgment and have certain limitations. FER systems can incessantly and objectively check patient facial expression

ISSN 2959-6157

changes, slight emotional fluctuations, provide doctors with a quantitative, accurate diagnostic basis [1, 2]. Help make a personalized treatment plan and improve diagnostic accuracy and treatment effectiveness.

In places of public gathering like train stations and airports, FER can also be combined. Do incessantly follow and target key personnel, improve security protection ability, and protect social stability. In the intelligent transportation sector, FER tech can obviously improve the transportation security protection level. It can be used on a driver status monitoring system, by identifying drivers' tiredness, anger, distraction, and other emotions, to send early warning signals promptly [3] remind drivers to rest, change status, and effectively prevent traffic events from happening.

In the early period, FER mainly depended on traditional machine learning algorithms, such as Support Vector Machine (SVM) and Adaptive Noise Reduction Algorithms (AdaBoost). When using these methods to perform facial expression recognition, it is very dependent on artificial design and feature extraction. For example, commonly used local binary patterns (LBP) compare the gray values of center pixels and neighboring pixels to produce a binary encoding. It describes local texture features of images and catches the texture variation of facial expressions. Histogram of Oriented Gradient (HOG) is used to calculate the HOG of the image‘s local areas, characterize the target shape, and obtain contour information. For muscle movement, it has a certain ability to portray facial contouring of facial expression. With the development of machine learning, FER has seen some big changes. Deep learning, through powerful automatic feature learning ability, can automatically extract high-level and abstract facial expressions from a large amount of data and obviously improve facial expression recognition accuracy. Among them, CNN dominates FER research.

With the development of machine learning, FER has seen some big changes. Deep learning, through its powerful automatic feature learning ability, can automatically extract high-level and abstract facial expressions from a large amount of data and obviously improve facial expression recognition accuracy. Among them, CNN dominates FER research.

The VGG model is a classic representative of the deep learning field. Take VGG16 as an example, it constructs a 16-layer deep network structure. This deep structure can effectively learn features from low-level edges and textures to high-level semantic features, such as facial expressions, eye squinting, corners of the mouth upward or downward, these facial muscle movement patterns, and change in form. In FER, many researchers use the VGG16 model pre-trained on large-scale image datasets as a feature extractor, extract the activation values before the fully connected layer as high-level feature representations of facial images, and then put them in a new classifier to perform facial expression category learning. Some other researchers involve further fine-tuning of the VGG16 model; thawing and retraining top layers can make parameters better adapt to specific task requirements for facial expression recognition and achieve higher accuracy than pure feature extraction.

Except for VGG, other typical CNN models are also widely used for inference. As the first in the ImageNet competition to emerge deep learning models, AlexNet successfully shows that deep learning has huge potential in image recognition missions and provides a new approach to FER research. ResNet introduces the concept of residual connections, breaking through the challenges of vanishing gradients or exploding gradients incurred by increased depth; it can construct over a hundred very deep network structures, and in many visual missions, can get the best performance at that time.

In recent years, Generative Adversarial Networks (GAN) and other emerging technologies have also been introduced into the field of FER. GAN is composed of a generator and a discriminator. The generator can learn to generate similar data to real facial expression images and effectively increase training in data diversity. It has important significance to address the lack of training data, data imbalance in FER, improve the model‘s generalization ability, and enable facial expressions to have better performance when facing complex real environments.

With deep learning models in FER missions becoming more and more complex, the lack of explanatory power has become a key barrier to the implementation of technology. In medical diagnosis and judicial assistance, these fields need high transparency, so a model that just outputs "anger" or "fear" classification results is not enough; researchers and users need to understand the basis on which the model makes judgments. For example, is it based on the corners of the mouth being pulled down or the eyebrows being furrowed?

This review aims to comprehensively and systematically analyze existing facial expression recognition methods. It summarizes different methods' core concepts, key technologies, advantages, and disadvantages, then compares and analyzes to figure out the applicability and development potential of different methods. In addition, it discusses challenges and future directions of development in the FER field, provides references to researchers, and promotes the next application and development of FER technology.

2 Methodology

FER methods always revolve around "feature extraction" and "Classification decision". Traditional machine learning methods depend on artificially designed features and statistical learning models, and deep learning methods use neural networks to enable end-to-end feature learning and classification. This chapter systematically describes the technical details of six major methods, including three classical machine learning methods: SVM, k-Nearest Neighbor (KNN), decision tree, and three deep learning methods: CNN, Recurrent Neural Network (RNN, and Long Short-Term Memory Network (LSTM), GAN. It lays the foundation for the next performance comparison and analysis.

2.1 Traditional Machine Learning Methods

Traditional machine learning methods are: artificial feature extraction to feature dimension reduction and to classifier training. Its performance highly depends on the quality of feature engineering, and it is applicable to scenes of data volume have limited and computing resources are constrained.

Support Vector Machine

The definition of SVM is to find the best classification hyperplane that maximizes the distance between samples of different categories. For non-linear partitionable problems, SVM uses nuclear functions to achieve linear separability by mapping low-dimensional features to a high-dimensional space.

The applications based on SVM are as follows: 1) feature input: the method normally uses artificial features such as LBP, HOG, or their integrated features. 2) nuclear function selection: because of the non-linear characteristics of expression features, the RBF kernel is most widely used in FER. It can effectively capture the complex relationship between facial texture and outline. 3) parameter optimization: use cross-validation to adjust the penalty coefficient (C) and kernel parameters (γ) to balance model complexity and generalization ability.

The advantages of SVM-based methods are strong generalization ability in small-sample scenarios, Insensitivity to high-dimensional features, and anti-overfitting capability better than classic classifiers. On the other hand, the disadvantages are dependent on the quality of artificial features, the exponential increase of computational complexity with sample size, and difficulty in processing large-scale FER datasets [4, 5].

k-Nearest Neighbor

The definition of KNN is simple: similar samples are placed closer together. By calculating the distance between an unlabeled sample and all samples in the training set, select the nearest samples and determine their category through majority voting.

The applications based on KNN are as follows: 1) pre-processing features: features such as LBPs and Gabor filters need to be normalized to avoid different dimensions in distance calculation. 2) K value selection: In FER, K is usually an odd number ranging from 3 to 11, and cross-validation is used to balance classification stability and noise sensitivity. 3) distance metric: Euclidean distance is used to balance stability of classification and sensitivity to noise; Manhattan distance shows greater stability in discrete features.

The advantages of KNN methods are that the model is simple and easy to achieve, it doesn't require a training process, and the dataset can be updated dynamically. On the other hand, the disadvantages are high computational cost, sensitivity to non-balanced data, and poor robustness to interferences such as shading and changes in lighting [6, 7].

Decision Tree

The definition of a decision tree is recursively partitioning of the feature space to construct tree models; every internal node represents a feature detection, and leaf nodes represent expression categories. The basis for partitioning is typically information gain or the Gini index, and the preference to choose that can mostly decrease data uncertainty.

The applications based on decision trees are as follows: 1) selection of features: they mostly employ facial key point features or geometric features, which are physically well-defined. 2) pruning strategy: by constraining the depth of the tree and removing redundant branches to avoid overfitting, and to improve the model's generalization ability. 3) integrated extension: the performance of a single decision tree has limitations; in practical applications, random forests and other integrated approaches are commonly used. Let many trees vote to increase accuracy in FER.

The advantages of methods based on decision tree are high explainability, low sensitivity to missing values, and the ability to process non-linear features. On the other hand, the disadvantages are that one decision tree is easy to overfit and has low processing efficiency for high-dimensional features. It has poor performance in recognizing subtle differences in expression.

ISSN 2959-6157

2.2 Deep learning methods

Deep learning methods use multi-layer neural networks to automatically learn features, to get the end-to-end process of "from original image to high-level features and to classification results". Its performance on large-scale datasets is far beyond that of traditional methods [4].

Convolutional Neural Network

The definition of CNN is to construct layer-based feature representation through local feature extraction, reduce dimensionality, translation invariance and classification decision. Key mechanisms include focusing only on specific areas of the image, reducing parameters and layer-wise feature learning.

The applications based on CNN are as follows: 1) classic structures: VGG16 uses 13 convolution layers and 3 fully connected layers to extract fine-grained facial expression features, which is applicable to FER for static images. ResNet adopts Residual Blocks to solve the gradient disappearance problem of deep works, and its 152-layer version achieves an accuracy of 85.6% on the FER-Plis dataset. 2) improve strategy: attention mechanism uses channel weight allocation to enhance key facial expression features. Transfer learning: it is based on pre-trained weights from ImageNet to initialization the network and freezing the bottom convolutional layers to speed up convergence and enhance the performance of small-sample FER.

The advantages of CNN methods are automatic learning of high-level features, strong characterization capability for complex expressions, and it can achieve an accuracy of 75%-90% on large-scale datasets [7]. On the other hand, the disadvantages are that they need large amounts of annotated data, have high computational cost and poor model explainability.

RNN and LSTM

The definition of a Recurrent neural network (RNN) is that it preserves timing information through hidden layer states and applies to solving dynamic facial expressions in video sequences. But it has the challenge of transmitting distant temporal features, and the long short-term memory network (LSTM) uses input gates, forget gates, and output gates to control information flow, effectively addressing this shortcoming.

The applications based on RNN/LSTM are as follows: 1) input form: it converts the video frame sequence into a CNN feature sequence and inputs them into the LSTM. 2) Time series modeling: it uses LSTM to catch dynamic changes of expression and output sequence-level emotion types. 3) mixed architecture: they usually use a "CNN+LSTM" combination, CNN extracts single-frame spatial features, and LSTM models interframe timing relationships. On FER video datasets, they achieve 10%-15% higher accuracy compared to static methods.

The advantages of the RNN methods are that they can catch the temporal dependence of dynamic expressions and apply to FER videos. On the other hand, the disadvantages are that the training process is prone to gradient explosion or disappearance; they are sensitive to noise frames; and their computational efficiency is low when processing long sequences.

Generative Adversarial Network

GAN is composed of a generator and a discriminator. It uses adversarial training to optimize: the generator generates realistic facial expression samples; the discriminator discriminates between real samples and generated samples and finally reaches a Nash equilibrium.

The applications based on GAN are as follows: 1) for rare categories like "disgust" and "fear" in FER datasets, the generator can generate diverse samples to mitigate category imbalance. 2) Feature learning: It uses the adversarial process of GAN to learn robust expression features. 3) noise reduction and enhancement: it uses GAN to fix occluded or blurred facial images and improve FER performance in complex scenes.

The advantages of GAN methods are that they can generate realistic samples to expand data and enhance the robustness of models against interference. On the other hand, the disadvantages are unstable training, high computational cost, and it is primarily used as an auxiliary tool in FER

3 Comparison

Method selection in the FER field directly impacts system performance and applicable scenarios. This chapter is based on the technology analysis of the 6 main methods in Chapter Two. This chapter conducts a comparison from both quantitative and qualitative dimensions: first, it conducts a horizontal comparison of the core differences among methods of the same type, then it conducts a longitudinal analysis of the fundamental differences between traditional machine learning methods and deep learning methods, and finally provides a basis for the decision of FER methods selection. The comparison and analysis will be carried out around these core dimensions:

· Performance metrics: The analysis focuses on accuracy, combined with a confusion matrix.

· Robustness: it evaluates the model's ability to assess disturbances such as changes in lighting, pose deviations, and occlusions.

· Data Requirements: It covers sample counts, annotation cost, and the model's sensitivity to data balance.

In addition, the Computational complexity is quantified by model parameter count and inference time. The experimental data is mainly obtained from the publicly available datasets, which include a rich collection of facial expression samples.

3.1 Traditional Machine Learning Model Comparison

Traditional machine learning methods all depend on artificial features but have obvious differences in classification mechanisms and performance. Table 1 summarizes the core comparison results for SVM, k-NN, and the decision trees in the FER task.

Table. 1 Comparison of CNN, LSTM and GAN core performance.

Comparison methods	SVM	KNN	Decision tree
Average accuracy rate	62% - 68%	55% - 62%	58% - 65%
Rare category recognition rate	50% - 55%	40% - 45%	45% - 50%
Deduction time	fast	slow	moderate
Robustness to light changes	moderate	low	moderate
Robustness to posture changes	low	low	moderate

SVM has the best performance, its classification advantages in high-dimensional space are obvious, and it has a high recognition rate of common expressions; then is random forest, which is based on a decision tree, enhances stability; KNN accuracy is relatively low.

SVM accuracy is higher than that of KNN in the recognition rate of rare categories; its feature of maximizing category spacing reduces interference from majority categories to minority types.

Decision trees perform better in scenarios with posture change, because the sensitivity of their geometry features to posture changes is lower than that of texture features; KNN is most sensitive to posture and lighting changes.

SVM applies to small- and medium-scale datasets, scenarios with limited computational resources, and scenarios with certain accuracy requirements. KNN is suitable for scenarios where samples are frequently updated dynamically, but it needs to be used in conjunction with dimensionality reduction technology. Random forests are suitable for scenarios that require a certain level of explainability; they can use feature importance ranking to identify key decision bases.

3.2 Deep Learning Model Comparison

Deep learning methods use automatic feature learning to achieve a performance breakthrough, but different models each have their own focuses in the capture of space features, time-series modeling, and data adaptation, the details are presented in Table 2.

Table. 2 Comparison of CNN, LSTM and GAN core performance.

Comparison methods	CNN	LSTM	GAN
Average accuracy rate	82% - 88%	78% - 85%	
Rare category recognition rate	65% - 70%	70% - 75%	
Deduction time	moderate	slow	moderate
Robustness to complexity scenarios	high	high	very high

ResNet shows the best performance in static images for FER; it can capture subtle texture features. LSTM shows clear advantages in video sequences; it can use time-series modeling to differentiate real and fake expressions. And, by generating extra samples, GAN can significantly improve CNN's recognition rate, but the training time will soar.

GAN achieves the best improvement in CNN performance in complex scenarios, simulating sample generation to enhance interference, which improves the model's adaptability; LSTM, through time-sequence coherence filtering for single-frame noise, and possesses good robustness.

ResNet and other CNNs are applicable for FER missions for static images, and researchers can select models of varying complexity based on requirements. LSTM can be used for videos, and it needs to be combined with CNN to extract frame features and temporal features. GAN is suitable for scenarios where the model needs high robustness,

ISSN 2959-6157

but it needs sufficient computational resources to support training.

3.4 Traditional Machine Learning and Deep Learning Whole Comparison

Traditional machine learning and deep learning are not a substitute relationship in FER missions; they are technical methods suitable for different scenarios. Table 3 conducts a horizontal comparison in terms of essential characteristics, then combines real application needs to provide select advice.

Table. 3 Comparing Traditional machine learning and Deep Learning Core Performance

Comparison methods	Traditional machine learning methods	Deep learning methods
Best accuracy upper limit	65% - 70%	85% - 90%
Data quality	high	low
Explainable	high	low
Hardware requirements	low	high
Deployment costs	low	Middle-high

For small sample scenarios, first select SVM or random forest; deep learning is easy to overfit, and annotation costs are high [8]. SVM and lightweight CNN are more applicable; LSTM and GAN have slow inference time with high real-time requirements. When high accuracy is needed and data is sufficient, select GAN-Enhanced ResNet and combine it with XAI technology to improve explainability. For dynamic facial expression analysis, CNN and LSTM are the only options, as classic methods can't catch temporal features. KNN and decision trees can rapidly achieve low cost, fast deployment, sacrificing some accuracy for development efficiency.

4 Challenges and Future Trends in FER

Although FER technology has achieved obvious development with traditional machine learning and deep learning, the progress from laboratory to real scenarios still faces limitations in multiple dimensions, such as data, models, and application environments. Furthermore, with the increasing requirements for technologies, new research directions are also emerging. Data is the foundation of FER model training; currently, the shortcomings in data quality and diversity have become the primary barrier to constraining the model's generalization ability.

Currently, most major FER datasets are collected under laboratory-controlled conditions, where the samples have uniform lighting, fixed positions, and simple backgrounds; in real scenarios, expressions are usually accompanied by complex interference, many postures, and occlusions [9].

In many FER datasets, "happiness", "calm", and so on, common expression samples account for over 40%, and "disgust", "fear", these rare emotions account for less than 5%. This imbalance causes the model to tend towards the majority class; the recognition rate of rare expressions is generally below 50%. In addition, the annotation progress is subject to subjective bias; different annotators have differentiated standard differences in "surprised" and "fear". It may cause dataset label noise, further decreasing the robustness of the model.

Expressions have cultural specificity; for example, in East Asian populations, when expressing "sadness", their facial muscle movements are more restrained, and Westerners tend to be more outgoing. At the same time, Individuals also express the same facial expressions differently due to age, gender, and facial structure.

4.1 Challenges at the Model Level

Even though under sufficient data conditions, the contradictions between FER model performance, quality, and explainability are still hard to balance. Currently, model recognition ability is weak for dynamic expressions and micro expressions. In dynamic expressions, facial muscle movement is hard to capture through single-frame features; but subtle muscle changes in micro expressions exceed the resolution of feature extraction in the CNN model, and it leads to micro expressions recognition accuracy of only 60%-70%. In addition, the model lacks discernment for "fake emotion".

High-accuracy model parameters account for over 60 million, training needs GPU support for several days, and inference time reaches 10-20 milliseconds per frame, making it hard to deploy on embedded devices; however, the lightweight model has reduced its parameters to 3.5 million, inference time reduced to 2ms per frame, but accuracy compared with ResNet decreases by 8%-12%

The Black-box feature of deep learning models is making its decision-making process difficult to trace. In medical, judicial, and other fields that need a highly transparent field, only outputting “anger” or “sadness” classification results is far from enough. Doctors need to know whether the model determines emotion based on “furrowed brows” or “drooping corners of the mouth” to verify the reliability of the results. Although XAI technology is trying to analyze the FER model, currently, the tools are hard to catch the key features of micro emotions and can’t explain the interrelations of features. It leads to explainability, remaining at the superficial level of “heat map display of areas of interest”.

On the application level, FER technology still needs to face environmental adaptation and user acceptance issues. In real scenarios, emotions are usually co-expressed through facial expressions, voice, physiological signals, and other multimodal information. But different sampling frequencies of modal data and significant differences in feature dimensions make data alignment difficult; Then, there may be “conflicts” between modes, existing integration strategies are hard to efficiently process this conflict, and the accuracy improvement of multimodal FER is only 5%-10%.

In real-time FER scenarios, the model is required to process a frame within 100 milliseconds, but currently, models that handle high-resolution videos often exceed 200 milliseconds of inference latency; in addition, FER needs to collect facial images, involving user privacy, if data leakage occurs during storage or transmission. It may lead to identity theft and emotional exploitation, and this privacy risk decreases users’ acceptance of the technology.

4.2 Future Trends in Facial Expression Recognition

Combining with computer vision and cutting-edge AI technology, FER is developing toward the direction of “more robust, more efficient, more explainable, and more secure” development. The key to solving data problems is increasing data diversity and authenticity, and emerging technologies are shifting from “expanding data volume” to “optimizing data quality” [10].

Diffusion models can generate more realistic and diverse expression samples compared with GANS. They can target the synthesis of fear expressions in low-light environments and disgust expressions with face mask occlusion, in these realistic complexity scenarios. It can also control samples’ culture, age attribute, efficiently mitigate data domain shifts, and category imbalance issues.

Through domain adaptation algorithms, the lab data “knowledge” is taken to move to realistic scenarios and decrease the domain offset effect. For example, in the training processes, a “domain resolver” is cited to let the model learn emotion classification and “distinguish whether data originates from laboratory settings or real-world scenarios.”, and finally make the model ignore scenario differences and only focus on emotion features. This method in cross-domain testing of FER-2013 and Real-World FER makes the accuracy decline rate from 32% to 15%.

5 Conclusion

This review of deep learning methods triggered FER to conduct systematic research, a comprehensive review of the FER field technology development process, core method system, performance differences, and current challenges. The outlook of the future development direction provides a clear technology reference structure for FER field researchers and practitioners.

In technology development and method system layers, FER technology has experienced a transition from traditional machine learning to learning key transition. The core step of a traditional machine learning method is “artificial features extraction to features dimensionality reduction and to the classifier training”, the SVM‘s advantage in high-dimensional space becomes the best choice, making it the best choice among classic methods. But this type of method highly depends on the quality of artificial features, and poor robustness makes it hard to be applicable for complex real scenarios. Deep learning methods use end-to-end feature learning to get performance breakthroughs: CNN in images FER shows powerful fine-grained feature capture capability; LSTM combines with CNN to construct a “space-sequences” mixed structure, efficiently resolves the dynamic expressions recognition problem; GAN, although not directly as the classifier, but it generates diverse samples, can effectively alleviate digital imbalance and soar the model’s robustness.

Using multiple dimensions comparison analysis, it can be found that these two methods are not a substitute relationship but a complementary relationship in scenarios. Traditional machine learning methods in small samples, low-resource, and requiring explainability scenarios still have practical value; deep learning method is mainly used in large-scale data, high-accuracy scenarios dominate, especially GAN-enhanced CNN models in complex scenes, robustness is best, but they need to bear higher computational costs and data demands.

In current FER from laboratory to real application, there are still three core challenges: in data layers, lab datasets, and „domain shift“ in real-world scenarios, the imbalance

ISSN 2959-6157

of expression types and individual differences leads to insufficient generalization ability of the model. In the model layer, the dynamic expressions and micro expressions recognition accuracy is low, and performance and efficiency are hard to balance. And the lack of explainability caused by the "black box" nature of deep learning constrains the application of the technique in key sectors such as healthcare and the judiciary. In the application layer, multimodal data alignment, insufficient real-time capability, and privacy leakage risks reduce the practical adaptability and users' acceptability of techniques.

References

[1] Cheng Qiyun, Sun Caixin, Zhang Xiaoxing, et al. Short-Term load forecasting model and method for power system based on complementation of neural network and fuzzy logic. Transactions of China Electrotechnical Society, 2004, 19(10): 53-58.

[2] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[3] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[4] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[5] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.

[6] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[7] SHI Biao, LI Yu Xia, YU Xhua, YAN Wang. Short-term load forecasting based on modified particle swarm optimizer and fuzzy neural network model. Systems Engineering-Theory and Practice, 2010, 30(1): 158-160.

[8] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[9] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[10] Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.