Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluatio
DOI: 10.54254/2755-2721/2025.PO24884

Explore the Facial Expression Recognition in Different Complex Environment

Haotian Lu[1*], Jiayi Xu[2], Haoteng Zheng[3]

[1]*School of Shanghai Liaoyuan Bilingual, Shanghai, China*
[2]*Shanghai Nanyang High School, Shanghai, China*
[3]*School of Computer Science and Engineering, Central South University, Changsha, China*
**Corresponding Author. Email: 8208220515@csu.edu.cn*

Abstract. Facial Expression Recognition (FER) holds extensive application value in fields such as human-computer interaction, intelligent security, and affective computing. However, it also faces challenges from complex conditions like illumination variations, occlusions, and head pose deviations, which cause significant performance degradation in traditional methods. This paper systematically analyzes recent advances in FER technologies addressing these issues and explores the application and innovation of deep learning methods in this domain. The analysis focuses on the following research directions: (1) Enhancement of recognition rate and robustness under illumination variations. Improved models incorporating transfer learning, activation function optimization, and anti-aliasing techniques have significantly boosted recognition accuracy. Approaches such as self-supervised learning and multi-feature aggregation demonstrate strong robustness and adaptability. (2) Feature compensation in occluded scenarios. Effective representation and facial feature-based methods are employed to restore occluded regions, improving recognition efficiency while fully leveraging the spatial structure of occlusions. This enables better representation and reconstruction of original samples. (3) Multi-pose facial expression recognition. By utilizing 3D deformable models for pose estimation and designing pose-invariant feature extraction networks, the challenges of expression recognition under large-angle head rotations are effectively addressed. This paper outlines the limitations of existing approaches and provides an outlook on potential future application scenarios and directions for technological breakthroughs.

Keywords: Facial Expression Recognition, Unconstrained Environments, Deep Learning, Self-Supervised Learning, Computer Vision

1. Introduction

Human facial expression is an important carrier of emotional expression, and studies have shown that in face-to-face communication, about 55% of the information is conveyed through facial expression. Nowadays, facial expression recognition has a wide range of applications in human-computer interaction(HCI), intelligent security, affective computing, driver monitoring, medical and

DOI: 10.54254/2755-2721/2025.PO24884

healthcare, etc. Therefore, accurate facial expression recognition is of great significance for realizing natural human-computer interaction and enhancing the ability of intelligent systems to understand emotions.

Currently, traditional facial expression recognition methods are susceptible to complex environments such as lighting variations, occlusion problems, and pose changes, which lead to significant degradation in recognition effectiveness. Ali et al. used IMF1 + IMF2 + IMF3 combination reconstructed by EMD (Empirical Modal Decomposition) to simulate JAFFE expression images at different luminance levels (30%, 40%, 50%, 60%, and 70%), and the recognition accuracy reaches 99.06% for the normal luminance condition, but the accuracy of the recognition gradually decreases with the decrease of luminance [1]. Rao et al. evaluated the recognition accuracy under different conditions, under unobstructed conditions, the accuracy of recognition can reach 99%, however, with sunglasses and scarf, the accuracy decreases to 81.6% and 86.6%, respectively [2]. Kuruvayil et al. tested the accuracy under complex backgrounds, the graphs existed with simultaneous variations of pose, occlusion, and illumination, and the accuracy was significantly reduced at multiple angles by 5-15% [3].

The development of deep learning techniques and improved conditions have gradually compensated for the limitations of traditional facial expression recognition methods. Convolutional neural network and Transformer-based models have achieved high recognition accuracy in some public datasets. Some multimodal fusion methods and dynamic sequence analysis have further improved the performance of facial expression recognition, and some lightweight models and edge computing technologies have made real-time expression recognition possible on mobile. However, the recognition effect under complex conditions is still a non-negligible problem, and the complex environment will still have an impact on the accuracy of the result, so improving the robustness under different environments is still a core challenge in the current field, and the development of environmentally robust facial expression recognition technology has important research value.

In this paper, we will introduce and analyze the current mainstream algorithmic models and key technologies for these environments for three types of complex environments, focusing on analyzing the strengths and weaknesses of each type of method and algorithmic performance, proposing the corresponding future practical applications and solutions, and exploring the direction of future breakthroughs.

2. Methods of comparison and analysis

In this paper, we focus on facial emotion analysis in three types of complex environments, namely, light-changing environments, occlusion problems, and complex environmental contexts, and compare and analyze the recognition effects with advantages and disadvantages in each case.

2.1. Under illumination variation

Light changes will lead to changes in the contrast and brightness of the image, affecting the expression of some edge features, and some regions are too bright or too dark will lead to the absence of some key information, so it is necessary to reduce the impact of light changes on the recognition effect in facial expression recognition. The most mainstream facial expression recognition method is the deep convolutional neural network (CNN), mainly due to its extraction of local details and sensitivity to expression changes, suitable for big data and structural changes. And for this method, researchers have made many innovative programs.

DOI: 10.54254/2755-2721/2025.PO24884

2.1.1. CNN approach

Firstly, a deep stacked convolutional self-coder (SCAE) is utilized to extract feature vectors that are insensitive to illumination changes by greedily and unsupervised training the pre-trained network layer by layer. Together with the variant ReLU activation function, the lighting effect can be adaptively adjusted when the brightness of the input image is abnormal.

The AlexNet and ResNet-101 models pre-trained on ImageNet can also be utilized for migration learning. Especially, ResNet-101 solves the deep network degradation problem by residual connection. The researchers validated the effectiveness of the illumination invariance approach by pre-training SCAE on Multi-PIE and fine-tuning the CNNs on KDEF and CK+, improving the classification accuracy by about 4% (KDEF, 95.58%) and 8% (CK+, 94.90%), respectively, and also achieving 100% accuracy for the four categories of emotions (disgust, happy, neutral, and cheerful) on CK+, superior to the accuracy of previous studies [4,5].

As for convolutional networks, in addition to SCAE, researchers have introduced an Anti-Aliasing Deep Convolutional Network (AA-DCN), which improves the FER performance through anti-aliasing techniques; Multi-task Cascaded Convolutional Neural Network (MTCNN): to achieve accurate face and key point detection, combined with DeepNet for feature extraction and emotion classification. The combination of ReLU, batch normalization, and Leaky ReLU techniques effectively mitigates the gradient vanishing problem [6].

By testing on CK+, JAFFE, and RAF-DB datasets, the researchers demonstrated that the AA-DCN model significantly improves the FER accuracy through the MaxBlurPool layer by 99.26% (CK+, with a 0.94% improvement), 98% (JAFFE, with a 3% improvement), and 82% (RAF-DB, with a 6% improvement), which is superior to the DCN and classical CNN models.

2.1.2. Innovative approach

In addition to CNN, there is a combination of using virtual and self-supervised comparative learning, which focuses on designing augmentation methods for perturbations such as lighting and occlusion with the dataset to improve model robustness. Then the vit model is applied for processing to get better results [7-9].

The researchers also proposed an innovative approach based on multi-feature aggregation, utilizing both Diverse Feature Extraction Module (DFEM) and the Hierarchical Attention Module (HAM). Complementary feature aggregation is achieved by taking into account local and global contextual features, high and low-level features, and gradient features that are robust to illumination. The discriminative features of critical facial regions are enhanced progressively to suppress irrelevant region interference. The method significantly improves the model's ability to capture spatial details and semantic information, and outperforms existing methods on several authoritative datasets, e.g., in the RAF-DB dataset, it achieves an accuracy of 91.23%, which is better than the 88.14% of SCN and the 87.03% of RAN. However, the DFEM and HAM modules increase the computational overhead, which may affect the real-time performance and high-resolution image processing efficiency [10].

2.1.3. Comparison

These methods have significant improvements in several datasets, and this section presents a partial display of the results of these methods in different datasets, each of which preprocesses the data and

DOI: 10.54254/2755-2721/2025.PO24884

uses 80% of the data for training and 20% for validation. Table 1 shows the specific results of each method.

Table 1: Accuracy of each method in the dataset

Datasets	Network Type	Network Size	Data Group	Performance
AffectNet-7 [9]	ViT + MoCo v3	12-layer Transformer	80-20 split	7 classes: 64.94
AffectNet-7 [8]	CNN with Attention	69.63M	80-20 split	7 classes: 65.74
AffectNet-8 [9]	ViT + MoCo v3	12-layer Transformer	80-20 split	8 classes: 60.63
CAER-S [8]	CNN with Attention	69.63M	80-20 split	7 classes: 91.16
CK+ [4]	CNN (SCAE)	5 conv layers	70-30 split	6 classes: 94.90
CK+ [6]	Deep CNN / Deep CNN + MaxBlurPool	3 stages	80-20 split	7 classes: 98.32 (DCN), 99.26 (AA-DCN)
CK+ [7]	DeepNet + Viola-Jones / MTCNN	Multi-layer	80-20 split	7 classes: 93.4 (Viola-Jones), 96.1 (MTCNN)
FER 2013	DeepNet + Viola-Jones / MTCNN	Multi-layer	80-20 split	7 classes: 83.6 (Viola-Jones), 84.5 (MTCNN)

2.2. In obstructed environments

Due to the inability to intuitively detect facial expressions and emotions when there are obstructions on the face, various methods need to be used to repair and guess the location of the obstruction as much as possible. In the presence of occlusion, one can attempt to repair the occluded position based on the characteristics of the character, in order to roughly obtain the appearance without occlusion and accurately recognize facial expressions to analyze emotions.

Sparse representation can be used. The main idea of sparse representation is to represent high-dimensional images in low dimensional space, establish independent subspaces for occluded areas, use existing dictionary atoms to represent occlusions, and then recognize occluded face images [11]. Sparse representation classification is achieved by using L1 norm regression to obtain the sparse representation coefficients of occluded face images in a dictionary and then reconstructing the occluded face images using the sparse representation coefficients to improve the performance of occluded face recognition [11]. This method has low complexity but cannot avoid errors.

The Embedding Module is also used to restore occluded facial features and suppress noise information in the features. It converts the descriptors extracted by the Proposal Module into similarity descriptors using the Locally Linear Embedding algorithm. Subsequently, the K-Nearest Neighbor algorithm was used to find matches in a feature pool that utilized a large number of unobstructed and occluded faces. Finally, by using the Verification Module, the repaired facial features can be utilized for facial region verification, fine-tuning the facial position and scale [12]. By setting reasonable anchors or large receptive fields, we can implicitly learn the faces in occluded areas. To reduce false recall, we can consider using segmentation or attention mechanisms to process images and obtain more accurate images. This method is more accurate compared to sparse

DOI: 10.54254/2755-2721/2025.PO24884

representation, but has a higher complexity. Due to the inability to intuitively detect facial expressions and emotions when there are obstructions on the face, various methods need to be used to repair and guess the location of the obstruction as much as possible. In the presence of occlusion, one can attempt to repair the occluded position based on the characteristics of the character, in order to roughly obtain the appearance without occlusion and accurately recognize facial expressions to analyze emotions.

Sparse representation can be used. The main idea of sparse representation is to represent high-dimensional images in low dimensional space, establish independent subspaces for occluded areas, use existing dictionary atoms to represent occlusions, and then recognize occluded face images. Sparse representation classification is achieved by using L1 norm regression to obtain the sparse representation coefficients of occluded face images in a dictionary, and then reconstructing the occluded face images using the sparse representation coefficients to improve the performance of occluded face recognition [13]. This method has low complexity but cannot avoid errors.

The Embedding Module is also used to restore occluded facial features and suppress noise information in the features. It converts the descriptors extracted by the Proposal Module into similarity descriptors using the Locally Linear Embedding algorithm. Subsequently, the K-Nearest Neighbor algorithm was used to find matches in a feature pool that utilized a large number of unobstructed and occluded faces. Finally, by using the Verification Module, the repaired facial features can be utilized for facial region verification, fine-tuning the facial position and scale. By setting reasonable anchors or large receptive fields, we can implicitly learn the faces in occluded areas. To reduce false recall, we can consider using segmentation or attention mechanisms to process images and obtain more accurate images. This method is more accurate compared to sparse representation but has a higher complexity (Table 2).

Table 2: The advantages and disadvantages of this method

Approach	Advantages	Disadvantages
Methods of sparse representation	Using a small number of samples to represent existing samples more compactly can improve the efficiency of successful recognition	Neglecting the spatial structure and continuity of occlusion
Processing of occlusion dictionary	Fully utilize the occluded spatial structure to better represent the original samples	The occlusion dictionary is single and cannot meet the diversity of occlusion types
Error coding of occlusion	Pay attention to the occlusion of one's own content, combine the occlusion structure with error coding, and improve recognition accuracy	Relying on a large number of training samples results in lower recognition efficiency under small sample conditions

2.3. Under unconstrained environments

In addition to illumination variations and occlusions, head pose variation represents another significant challenge for facial expression recognition in wild. The head pose is typically described using Euler angles, which consist of three components: pitch, yaw, and roll. The head pose may lead to partial facial occlusion, preventing key expression features (e.g., furrowed brows or raised mouth corners) from being fully visible. It can also distort facial geometry, altering the appearance of facial features. These variations make it difficult for algorithms to capture relevant expression features effectively, ultimately compromising the accuracy of emotion recognition. Consequently, the

DOI: 10.54254/2755-2721/2025.PO24884

integration of Head Pose Estimation (HPE) has emerged as a critical technique for enhancing the robustness of affective recognition systems.

To address head pose variations, algorithms typically incorporate a head pose estimation layer to process input images. This layer estimates the head pose and performs head image alignment (e.g., generating frontal-facing images) to facilitate subsequent facial feature extraction and emotion recognition. Depending on whether they rely on facial landmarks, head pose estimation algorithms are generally categorized into two types. (1) Landmark-based methods: These approaches leverage geometric relationships and relative positions of facial landmarks to estimate head pose. (2) Landmark-free methods: These techniques typically employ deep learning to extract facial features directly and infer head pose without explicit landmarks detection.

Benefiting from the rapid development of neural network technology, deep learning-based landmark detection methods have significantly improved the precision of landmarks localization, thereby enhancing the accuracy of head pose estimation.

QDLBP-Net employs 3D-FAN to localize 68 facial 3D landmarks and computes head pose based on these landmarks [14]. This landmark localization method adopts a fully convolutional neural network architecture that directly regresses heatmaps of facial landmarks through stacked Hourglass modules [15]. For each landmark, it generates a 2D Gaussian heatmap where the peak position corresponds to the landmark coordinates. The approach further incorporates hierarchical, parallel and multi-scale structures by combining feature maps at different resolutions, significantly improving robustness against challenging scenarios such as large poses and occlusions.

For facial samples captured in the wild, perspective distortion frequently occurs due to imperfect alignment between faces and cameras. AccuHPE introduces a preprocessing layer that corrects perspective distortion in raw samples [16]. The approach employs a lightweight CNN network to perform head pose estimation on the rectified images, and feedback the estimation result to the camera coordinate system, thereby achieving high-precision head pose estimation.

In traditional approaches, landmark-based methods heavily rely on accurate landmark detection, demonstrating limited capability in scenarios involving extreme poses, complex illumination, and occlusions. Conversely, landmark-free methods discretize continuous head movements into finite, categorical outputs, resulting in pose features lacking rotational information and being susceptible to ambiguities in pose labeling.

The diffusion models for head pose estimation establishes a reverse diffusion process on the Special Orthogonal Group in 3D (SO(3)) manifold, progressively denoises and iteratively refines the mapping to enhance estimation accuracy and roubustness [16]. The method defines a non-uniform Markov chain that quantifies rotation distributions at arbitrary timesteps using isotropic Gaussian distributions, guiding the diffusion process through rotational distributions on SO(3) for effective sampling. A cycle-consistency module is incorporated as an intermediate constraint, where K-nearest neighbors are selected to construct candidate sets, thereby enhancing feature representation diversity and richness while capturing high-level semantic information [16].

Deep learning-based head pose estimation heavily relies on large-scale accurately labeled samples. Current datasets for unconstrained head pose estimation research face significant limitations - they either contain substantial amounts of non-realistic synthetic samples or are constrained to small-scale natural images with inaccurate annotation. Consequently, fully supervised solutions remain restricted due to the dependence on extensive accurate labeling.

Semi-Supervised Unconstrained Head Pose Estimation (SemiUHPE) leverages large amounts of readily available unlabeled head images to address dataset limitations [17]. The method employs semi-supervised rotation regression techniques to adapt to both error sensitivity and label scarcity in

DOI: 10.54254/2755-2721/2025.PO24884

unconstrained head pose estimation [17]. Building upon the observation that aspect-ratio-invariant cropping outperforms landmark-based affine alignment, it introduces a dynamic entropy-based filtering approach and two novel head orientation augmentation methods [17].

Head pose estimation provides crucial spatial context and viewpoint robustness for facial express recognition. Future research should focus on integrating unsupervised learning, multimodal fusion, and lightweight architectures to advance its practical deployment in dynamic real-world scenarios.

3. Change and challenge

In today's complex environment, emotional perception still faces many challenges. Firstly, at the technical level, there is a balance between real-time performance and resource optimization. Facial expression recognition technology requires real-time processing and analysis of user expressions in complex environments, which poses a huge challenge to hardware computing power and software processing efficiency. At the same time, the hardware cost and energy consumption issues of smart home devices also limit the popularization of technology;

At the data level, the effectiveness of model training is closely related to the dataset. However, in complex environments, the amount of data in the dataset still accounts for a small proportion of the current data, and the imbalance of various emotions can also affect the accuracy of the final recognition effect. Therefore, it is necessary to increase the amount of data that can be collected in complex environments as much as possible to better and more accurately recognize.

At the user level, facial data belongs to highly sensitive information, which imposes strict requirements on the secure storage and transmission mechanism of data. The system needs to ensure that data is not abused or leaked, and comply with relevant laws and regulations. The popularization of technology also faces concerns from users about privacy, security, and technical reliability. In the promotion process, attention should be paid to users' psychological and socio-cultural factors, and efforts should be made to improve social acceptance. Users should be informed in advance of the privacy and other convenient needs that need to be disclosed before the application.

Overall, facial expression recognition in complex environments has a wide range of application markets. It can be applied to existing facial expression recognition fields to improve recognition performance, such as intelligent driving detection and medical health monitoring, achieving recognition accuracy in various complex environments. The optimal development is to prioritize its application in the fields of medical health and daily life. In addition, current technology can be transferred to other fields such as facial recognition and gender recognition to solve similar problems in these areas, which also has significant effects on the progress of these fields.

4. Conclusion

This paper provides a systematic analysis of recent advancements in FER technologies under unconstrained environments, including illumination variations, occlusions, and head pose deviations. Furthermore, it explores the applications and innovations of deep learning approaches in this field.

Under illumination variation, the introduction of improved models such as SCAE, AA-DCN, and MTCNN, combined with transfer learning, activation function optimization, and anti-aliasing techniques, has significantly enhanced recognition accuracy. Additionally, approaches integrating ViT with self-supervised learning and multi-feature aggregation methods have demonstrated superior robustness and adaptability. Collectively, these methods effectively mitigate illumination-induced interference and achieve performance surpassing traditional approaches across multiple

DOI: 10.54254/2755-2721/2025.PO24884

datasets. However, certain methods exhibit high computational complexity, indicating the need for further optimization to balance accuracy and efficiency.

In occlusions, the proposed approach employs sparse representation and an Embedding Module to reconstruct occluded facial regions. Subsequently, the Locally Linear Embedding algorithm is utilized for precise local restoration of occluded areas, followed by K-Nearest Neighbor matching to identify the most similar facial samples. A dedicated Verification Module then performs the final facial authentication. This framework not only enhances recognition efficiency but also effectively leverages the spatial structure of occlusions, enabling more accurate representation and reconstruction of original samples.

In facial expression-based emotion recognition, head pose estimation plays a critical role in achieving a comprehensive understanding of emotional expressions. A series of innovative approaches have emerged, continuously advancing the accuracy, robustness, and real-time performance of head pose estimation. Geometry-based methods utilize facial landmarks combined with models like 3D Morphable Models (3DMM) to fit 3D rotation matrices and compute pose parameters. Deep learning-based approaches employ various network architectures to regress pose angles from images. Representative models like FSA-Net, HopeNet, and HeadDiff have significantly improved the accuracy and robustness of head pose estimation. To accommodate real-time applications, lightweight networks such as MobileNet have been adapted for head pose estimation. Meanwhile, semi-supervised learning frameworks like SemiUHPE leverage unlabeled data to optimize pose representation while reducing annotation costs.

Looking ahead, facial expression recognition technology will continue to undergo profound innovations in algorithms, hardware, and software. The further employment of deep learning and multimodal fusion techniques will drive continuous improvements in both accuracy and efficiency. As the technology matures and market applications expand, facial expression recognition will be broadly adopted across various fields. In particular, this technology is expected to play a significant role in human-machine interaction and technology-driven healthcare, enhancing management efficiency and service quality.

Authors contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] H. Ali, W. K. Wan Ahmad, H. Muthusamy, and M. Elshaikh, "Illumination effects on facial expression recognition using empirical mode decomposition, " in Proc. ICAROB, vol. 29, pp. 653-660, Feb. 2024.

[2] K. P. Rao and M. V. P. C. S. Rao, "Illumination invariant facial expression recognition using convolutional neural networks, " Int. J. Recent Technol. Eng., vol. 8, no. 4, pp. 6140-6144, Nov. 2019.

[3] S. Kuruvayil and S. Palaniswamy, "Emotion recognition from facial images with simultaneous occlusion, pose and illumination variations using meta-learning, " Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 9, pp. 7271-7282, 2022.

[4] A. Ruiz-Garcia, V. Palade, M. Elshaw and I. Almakky, "Deep Learning for Illumination Invariant Facial Expression Recognition, " 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1-6, 2018.

[5] Karu Prasada Rao and M. V. P. C. S. Rao, "Illumination Invariant Facial Expression Recognition using Convolutional Neural Networks." International Journal of Recent Technology and Engineering (IJRTE) 8.4 (2019): 6140-6144.

[6] R.A. Elsheikh et al., "Improved facial emotion recognition model based on a novel deep convolutional structure." Scientific Reports 14 (2024): 29050.

[7] Sukhavasi SB et al., "Deep Neural Network Approach for Pose, Illumination, and Occlusion Invariant Driver Emotion Detection." International Journal of Environmental Research and Public Health 19.4 (2022): 2352.

DOI: 10.54254/2755-2721/2025.PO24884

[8] Huanjie Tao and Qianyue Duan, "Hierarchical attention network with progressive feature fusion for facial expression recognition." Neural Networks 170 (2024): 337-348.
[9] Cui Xinyu, He Chong, Zhao Hongke et al. "Facial expression recognition by integrating ViT and contrastive learning." Journal of Image and Graphics, 29.1 (2024): 123-133.
[10] Licai Sun et al., "MAE-DFER: Efficient Masked Autoencoder for Self-supervised Dynamic Facial Expression Recognition." Proceedings of the 31st ACM International Conference on Multimedia (2023).
[11] Yang M et al., "Fisher discrimination dictionary learning for sparse representation." IEEE International Conference on Computer Vision (2011): 543-547.
[12] Wang Huixing, Huang Bo, Gao Yongbin et al. "A review of some methods for face recognition with partial occlusion." Journal of Wuhan University (Science Edition) 66.5 (2020): 451-461.
[13] OU W H，LUAN X，GOU J PJ N al. Robust discriminative nonnegative dictionary learning for occluded face recognition ［J］. Pattern Recognition Letters，2018，107：41-49. DOI：10.1016/j.patrec.2017.07.006.
[14] Lianghai Jin et al., "Quaternion Deformable Local Binary Pattern and Pose-Correction Facial Decomposition for Color Facial Expression Recognition in the Wild." IEEE Transactions on Computational Social Systems 11.2 (2024).
[15] Bulat A and Tzimiropoulos G, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230, 000 3D facial landmarks)." IEEE International Conference on Computer Vision (2017): 1021-1030.
[16] Xiao Li et al., "Accurate Head Pose Estimation Using Image Rectification and a Lightweight Convolutional Neural Network." IEEE Transactions on Multimedia 25 (2023).
[17] Yaoxing Wang et al., "HeadDiff: Exploring Rotation Uncertainty With Diffusion Models for Head Pose Estimation." IEEE Transactions on Image Processing 33 (2024).