A Review of Facial Expression Recognition Issues, Challenges, and Future Research Direction

Yan Bowen, Azween Abdullah, Lorita Angeline, and S.H. Kok

School of Computer Science and Engineering, Taylor's University, Malaysia

Abstract
Facial expression recognition, a topical problem in the field of computer vision and pattern recognition, is a direct means of recognizing human emotions and behaviors. This paper first summarizes the datasets commonly used for expression recognition and their associated characteristics and presents traditional machine learning algorithms and their benefits and drawbacks from three key techniques of face expression; image pre-processing, feature extraction, and expression classification. Deep learning-oriented expression recognition methods and various algorithmic framework performances are also analyzed and compared. Finally, the current barriers to facial expression recognition and potential developments are highlighted.

Keywords:
Facial expression recognition; image pre-processing; feature extraction; machine learning; deep learning

1. Introduction

Although technological advancements in artificial intelligence have enabled people to manage highly complex problems, current human-computer interactions are yet to reach the ideal state of intelligence given the computer's lack of perception of the inner human world. Research by Mehrabia, a renowned psychologist, indicated the role of facial expressions in conveying the largest proportion of emotional information in the human emotional expression pathway [1]. Meanwhile, Ekman and Friesen's extensive cross-cultural study defined six basic expressions: happiness, fear, anger, sadness, surprise, and disgust. The authors then incorporated 'contempt' into the basic expressions and built the first face expression database encompassing thousands of distinctive expressions. A facial action coding system (FACS) encompassing over 40 facial action units [2] was proposed to perform broader and multi-category face expression portrayal and classification following the limitations underlying the basic expression emotion model. Thus, facial expression recognition has garnered much scholarly interest in understanding human beings' corresponding psychological states through computers.

The fundamental process framework of facial expression recognition, as depicted in Figure 1, implies a typical image acquisition, image pre-processing, feature extraction, and expression classification problem [3]. Notably, the algorithmic framework of facial expression recognition has transitioned from machine learning to deep learning algorithms after decades of development. Facial expression recognition with a higher recognition accuracy would alter the human-computer interaction level despite the initial application of facial expression recognition to distance education, business analysis, and auxiliary medicine domains.

This paper, which reviewed recent works on machine learning and deep learning in facial expression recognition, first summarized publicly-accessible and popular facial expression datasets. Meanwhile, the benefits and drawbacks of conventional machine learning and deep learning algorithms were presented in terms of image pre-processing, feature extraction, and feature classification. The challenges underpinning facial expression recognition and potential development trends are also presented in this article.

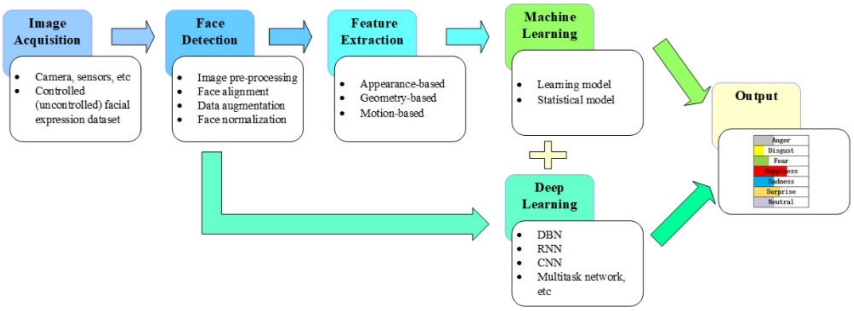

Fig.1 The general pipeline of deep facial expression recognition systems

Manuscript received January 5, 2023
Manuscript revised January 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.1.17

2. Common Facial Expression Dataset

Enriched and functional facial expression databases must be established to develop facial expression recognition technologies, as facial expression databases provide essential facial expression data for facial expression recognition. Present facial expression databases are categorized into (i) laboratory environment and (ii) real environment. Table 1 highlights the typical datasets in current expression-oriented study domains based on scale size, collection environment, and annotation situation.

Table 1: Basic Information on Common Facial Expression Dataset

Datasets	Data Sample	Subject	Conditions	Expression Distribution
JAFFE [4]	213 images	10	Lab (front-induce)	7 basic expressions
CK+ [5][6]	593 image sequences	123	Lab (front-induce-spontaneous)	7 basic expressions and AU
MMI [7]	740 images and 2900 videos	25	Lab (induce)	7 basic expressions
BU-3DFE [8]	2500 images	100	Lab (induce)	7 basic expressions
GavabDB [9]	549 images	61	Lab (induce)	7 basic expressions
SMIC [10]	164 image sequences	16	Lab (induce)	3 micro-expression types
CASME II [11]	246 image sequences	26	Lab (spontaneous)	7 micro-expression types
FER2013 [12]	35887 images	N/A	Web (induce-spontaneous)	7 basic expressions
RAF-DB [13]	29672 images	N/A	Web (induce-spontaneous)	7 basic and 12 compound expressions
SFEW2.0 [14]	1766 images	N/A	Movie (induce-spontaneous)	7 basic expressions

1) Lyons et al.'s [4] JAFFE dataset, which was built based on a laboratory setting, contains six basic and neutral expressions made by 10 Japanese women in the same context, for a total of seven expressions. Each expression encompassed 30 to 31 images for a total of 213 images following the small sample size. Notably, the images are of the same size with specific differences in light intensity.
2) Currently the most widely used dataset for FER system evaluation in an experimental setting, the CK+ dataset [5-6] contains 593 video sequences gathered from 123 experimenters. Each video sequence demonstrates the complete transition of facial expressions from neutral to peak, which lasts for different durations. Specifically, 327 video sequences were labeled with emotion in addition to a sample category of light contempt expressions relative to the six basic expressions in the CK dataset.
3) The MMI dataset [7] is similarly collected in a laboratory setting. Unlike CK+, the MMI sequences begin with a neutral expression and peak near the middle before returning to the neutral state. The 326 sequences involving 32 subjects contain 740 high-resolution still images and over 2900 videos, with 231 sequences labeled with six basic expressions. Essentially, the subjects have eyes and beards and are balanced in age and ethnic distribution.
4) The BU-3DFE dataset [8] provided by Binghamton University included 100 individuals, each of whom was captured in 25 images, including six basic expressions and one neutral counterpart. Each one is subsequently divided into four levels. Overall, 606 facial sequences and 2500 images were captured.
5) The GavabDB dataset [9] contains 549 3D images of 61 people (45 males and 16 females) ranging from 18 to 40 years old. Nine images were collected for each acquisition: two neutral frontals, two left-right full lateral, one head-up, one head-down, one smiling expression, one laughing expression, and one casual expression gesture image.
6) The SMIC dataset [10] created by Oulu University et al. is the first spontaneous micro-expression dataset that contains 164 micro-expression sequences recorded by 16 subjects: 10 males and six females. Specifically, the dataset constitutes three expression categories (positive, negative, and surprise), with positive expressions containing happiness and negative ones encompassing sadness, fear, and disgust.
7) The CASME II dataset [11] entails 246 micro-expression sequences that contain 26 individuals. All the micro-

Manuscript received January 5, 2023
Manuscript revised January 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.1.17

expression samples are also spontaneous with neutral beginnings and endings. Seven micro-expression types are presented as follows: happiness, disgust, surprise, repression, fear, and sadness, while others denote micro-subject-produced expressions with minimal distinctive features.

8) As a large Internet database automatically collected by the Google image search engine, the FER2013 dataset [12] contains 35887 face images that are divided into three parts: 28709 images as the training set data, 3589 as the validation set, and 3589 as the test set. The expression images were divided into six basic plus neutral expressions. Significant variances are identified between different images based on age and posture, which closely resemble the real world.
9) The RAF-DB dataset [13] developed by a research team at the Beijing University of Posts and Telecommunications contains almost 30,000 highly diverse facial images downloaded from the Internet. The database provides seven precise categories of basic expression labels (same as FER2013) and 12 composite expression labels for the sample. Meanwhile, the dataset is classified into basic and composite databases with the basic database divided into a training and test set. The test set constitutes 3068 images, whereas the training set encompasses 12,271 images.
10) The SFEW dataset [14] was created by selecting static frames from the AFEW database by computing keyframes based on facial point clustering. This dataset contains three parts: SFEW-Train, SFEW-Val (validation set), and SFEW-Test with 958, 436, and 372 samples, respectively. Each image was assigned to six basic and neutral expressions involving large pose head variations and a broad range of facial occlusions.

3. Facial Expression Recognition Based on Conventional Machine Learning Algorithms

Facial expression recognition based on traditional machine learning is primarily divided into three parts: face detection (the foundation), feature extraction (the core), and expression classification (the goal).

3.1 Face Detection

3.1.1 Image pre-processing

As factors involving illumination, pose, and occlusion significantly influences expression recognition accuracy, face images should be pre-processed with corresponding data to thoroughly eliminate the influence of interference factors on feature extraction. Traditional image pre-processing includes segmentation and statistical methods. For example, Hallinan et al. [16] categorized the human face into local regions and constructed an energy function to identify salient features with a variable model of the eyes and mouth for face detection. Phimoltares et al. [17] utilized the shape properties of the face for edge detection through the Canny operator and proposed a neuro-visual model for facial feature point-matching. Meanwhile, Wang et al. [18] who applied the statistical HMM to distinguish whether a face belongs to a facial region added a curvilinear wave transform with directionality for improved detection robustness.

3.1.2 Face alignment

Face alignment mainly depends on the input image to locate the face key points, including eye corners, eyebrows, chin, and nose. In Cootes et al.'s [19] ASM-based method, a shape model was first constructed from the face key points. A normalization method was subsequently used to obtain the average face model in the training dataset. Finally, the offset from the average face was employed for face key-point identification in the new image. One researcher [20] designed the IntraFace software following the improved AAM method to implement face localization with SDM and identify 49 accurate facial marker points by cascading the regression function to map the image appearance to key points, such as two eyes, a nose, a mouth, and two eyebrows.

3.1.3 Data augmentation

Random perturbation and pixel transformations denote common operations for traditional data augmentation. Pixel transformations include rotation, shift, tilt, scaling, adding noise points, contrast adjustment, and color dithering. Simard et al. [21] proposed increasing the number of samples by integrating these three spatial transformations by geometrically transforming the original image through rotations, panning, and tilting. Krizhevsky et al. [22] expanded the training set 2048 times by arbitrarily cropping fixed-size sub-samples on the original image and horizontally flipping each sub-sample. By incorporating random noise near the eyes of the original image with a two-dimensional Gaussian distribution, Lopes et al. [23] generated new two-eye positions and a rotation operation to horizontally position the new samples. Zavarez et al. [24] also applied a warping procedure to distort the image corners for training set expansion.

3.1.4 Face normalization

As the elicited face images depend on variations in illumination and head pose, both illumination and pose normalization were employed to rectify these concerns. Several scholars [25][26] recommended using both histogram equalization to improve the global contrast of the image for pre-processing, which is relatively effective when the background and foreground brightness are similar. Recent research has employed pose normalization to generate frontal views of facial expressions. Li et al. [27], who utilized normalization based on homomorphic filtering to omit illumination normalization, yielded good results For example,

Hassner et al. [28] first generated a 3D texture reference model common to all faces upon locating the face key points to estimate existing face components and synthesized the initial frontalized face by back-projecting each input face image to the reference coordinate system.

3.2 Feature Extractions

Feature extraction is a core component of the entire facial expression recognition system. As such, the extracted features need to fully represent the fundamentals of facial expressions, omit interference that is irrelevant to expressions (noise and lighting), process higher-latitude data, and perform feature dimensionality reduction and feature decomposition. A strong differentiation was identified between different expression categories for optimal outcomes. Nevertheless, image feature information in the traditional method is extracted manually, with primary reliance on researchers' experience. Table 2 presents the common classification.

3.2.1 Appearance-based features

Apparent features, such as Gabor, LBP, and Haar features ideally reflect the face texture structure. Following Lyons et al. [29], the Gabor wavelet method depends on local features, with the image processed using a filter with local relations to detect texture changes in multiple directions. Zhen [30] manually selected a local small region, extracted the Gabor wavelet coefficients on this region, and subsequently identified its mean value to be used as a texture feature. In terms of advantage, apparent features demonstrates lower effects on illumination changes, whereas the disadvantage lies in its computationally intensive nature and poor real-time performance.

Feng et al. [31], who generated pixel-point LBP features, formed statistical histograms on small bins and finally classified expressions on global statistical histograms. Furthermore, Guo et al. [32] proposed an improved LBP operator (FCL-LBP) based on Fisher's criterion and different dimensionality reduction methods for the existence of large data redundancy in the extracted features for improved recognition rates. The LBP algorithm could optimally represent local texture features by comparing the grayscale change relationship of images. Notwithstanding, the high-dimensional histogram generated by its algorithm renders its real-time performance poor and susceptible to noise interference.

Haar features, which denote high computing speed for image computation, were first proposed by Viola [29]. Satiyan et al. [33] employed Haar wavelet features for facial description in combination with both multiscale and statistical analysis for facial recognition. Furthermore, Min Hu et al. [34] recommended Haar features with histogram weighting to adequately describe the local features.

3.2.2 Geometry-based features

Face extraction methods for geometric features typically emphasize the positional relationships between feature points at the global level in terms of location, scale, and ratio information between one another. Pantic et al. [35] recommended a confirmatory coefficient approach that assigns coefficients to the extracted face feature points and determines the outcome based on the comparison between these coefficients. Both AAM and ASM algorithms could extract facial expression features, with AAM built on ASM to reflect the overall texture information of facial expression images. Regardless, this class of methods proves more suitable for coarse classification [19][20]. Following relevant research, improved face expression recognition requires integration with other methods. Cristinacce [36] combined PRFR with AAM to detect feature points on local edges, such as face features, while Saatci [37] cascaded AAM with an SVM classifier for high recognition rates.

3.2.3 Motion-based features

The aforementioned methods are ubiquitous in static face images. It is also deemed easier to judge the expression category in the presence of significant shifts. Nevertheless, much expression information requires extraction through the optical flow and image differencing methods following the temporal relationship of expression motion changes if the expressions are dynamic sequences. Ma [38] performed a differencing operation by integrating expression images with neutral ones and a discrete cosine transform on the different images to take the low-frequency part as the feature vector for expression recognition. Although the image differencing algorithm is simple in logic and small in operations, the disadvantage lies in requiring the expression images to be strictly aligned with one another and its susceptibility to interference. The optical flow method could highlight the face contour benefits and reflect face motion trends. For example, Wu et al. [39] recommended an improved instantaneous position velocity estimation method to overcome the drift problem of the optical flow algorithm and accelerate the convergence speed. Micro-expression keyframes were captured by Li et al. [40] based on the optical flow direction information entropy statistics. The direction information entropy of the image was obtained by counting the angle of optical flow change, which was analyzed to achieve the capture. In terms of drawbacks, this technique is sensitive to illumination, computationally intensive, and time-consuming.
The aforementioned feature extraction algorithms encounter a series of problems, such as dimensional catastrophe following high dimensionality despite extensive utilization. Thus, multiple feature extraction algorithms have been proposed and improved, with deep learning-oriented feature extraction gradually garnering scholarly interest.

Table 2: Traditional Machine Learning Feature Extraction Algorithms

Type	Author	Method	Contribution	Limitation
Appearance-based features	Lyons et al. [29]	Gabor (multi-orientation multi-resolution)	Smaller effect on light variation	High computational volume and poor real-time
	Zhen et al. [30]	Gabor (ratio-image)		
	Feng et al. [31]	LBP	Capable of representing the grayscale change relationship of the image	Poor real-time performance and susceptibility to noise interference
	Guo et al. [32]	FCL-LBP		
	Satiyan et al. [33]	Haar	Effectively describes light and dark changes	Disregarding the face shape and contour
	Hu et al. [34]	HCBP (Haar-Histogram)		
Geometry-based features	Pantic et al. [35]	Confirmability Factor	Focus on positional relationships between feature points at the global level	Suitable for coarse classification
	Cootes et al. [19]	ASM		
	Xiong et al. [20]	AAM		
	Cristinacce et al.[36]	AAM-PRFR	Focus on edge feature points	More subjective, complex operation, and sensitive to light conditions
	Saatci et al. [37]	AAM-SVM	Cascade with high recognition efficiency	
Motion-based features	Ma et al. [38]	Image Difference	Simple logic and minimal arithmetic	Images need to be aligned and are susceptible to interference
	Wu et al. [39]	Optical flow	Highlight the face contours and reflect face movement trends	Sensitive to light, computationally intensive, and time-consuming
	Li et al. [40]	Entropy of Oriented Optical Flow		

3.3 Facial Expression Classification

The category of expressions is determined by the classifier post-feature extraction. The appropriate classifier plays a pivotal role in classifying expressions, with the specific method decided based on the image category and divided into learning models [Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and AdaBoost] and statistical models [Naive Bayes (NB) and Hidden Markov Model (HMM)].

3.3.1 Machine learning model

Notably, SVM denotes a class of generalized linear classifiers that performs data segmentation by mapping non-separable linear data to a high-dimensional space with kernel functions. Huang et al. [41] derived the extracted facial geometric features through feature fusion and subsequently employed SVM for facial expression recognition. Several scholars proposed sparse SVM by improving the SVM itself [42], while other counterparts achieved good results with a combination of methods, such as integrating Gabor wavelets and SVM [43]. Although SVM could manage high-dimensional data, it is only suitable for processing small sample data.

The KNN is a more mature and simpler method. A given dataset seeks the nearest K samples each time for a new data instance. The new instance is categorized into this class if most of the K samples belong to a certain class. Chao [43], who utilized the nearest neighbor method to classify face expressions, indicated that the accurate classification rate of the nearest neighbor method relies on the number of samples to be classified. Xu et al. [44] integrated KNN with SVM, while Sun et al. [45] proposed

combining KNN with the region of interest (ROI), thus significantly enhancing the recognition algorithm robustness. This method reflects some disadvantages in terms of low training speed and accuracy following uneven sample distribution.

The main idea underpinning AdaBoost is to obtain the best weak classifier through continuous training on the same training set and synthesize several weak classifiers into one strong classifier upon satisfying a threshold number of iterations. Kandemir et al.'s [46] applied research employed Haar features to describe facial texture information and AdaBoost to perform key part screening in combination with geometric structure information to render comprehensive judgments. Regardless, AdaBoost is deemed more sensitive to outliers and noise and time-consuming.

3.3.2 Statistical model

As a mathematical statistical model based on the probabilistic relationship between state transfer and state observation sequences to estimate the implied parameters and develop a prediction model, HMM is generally utilized to resolve sequence-based problems. Lirong et al. [47] presented HMM as an expression classification method in facial expression recognition. A local binary model was first employed for the feature extraction of facial expressions, while the extracted features were embedded into the HMM model for classification. The integration of both approaches resulted in a higher recognition rate. Zhan et al. [48] recommended the FBMM algorithm following the HMM algorithm, which relaxed the conditional independence assumption of the traditional HMM counterpart for improved algorithm robustness and recognition rate. Devi et al. [49] further optimized HMM in terms of time complexity and parameters to parallelly process images and enhance classification efficiency.

The NB is a graphical network based on Bayesian formulation and probabilistic inference. See [50] proposed a probabilistic Bayesian classifier model to recognize faces in video sequences, match the representative images of video frames to subject classes based on the presented joint probability function, and conduct simulation experiments on two standard face video datasets for high recognition rates. The algorithm proves advantageous following its simplicity and insensitivity to missing data and disadvantageous as it requires independent expression features, which could be limiting.

4. Facial Expression Recognition Based on Deep Learning Algorithms

Relevant algorithms are extensively employed in facial expression recognition domains given the rapid development of deep learning technology. This technique, which could learn features with discriminative power, breaks the traditional fixed pattern of feature extraction, followed by pattern recognition in expression recognition. The method could also simultaneously perform feature extraction and expression classification. Researchers have improved expression recognition accuracy based on multiple perspectives, such as (i) from feature optimization to data enhancement and (ii) network structure changes to newly proposed classifiers.

4.1 Face Detection

4.1.1 Image pre-processing

Deep learning also requires data pre-processing to eliminate backgrounds and image differences in lighting and head pose that are irrelevant to facial expressions. As such factors affect the feature learning of deep networks, data pre-processing proves necessary before training. Traditional image pre-processing manually extracts features and omits backgrounds and non-facial regions. Ross et al. [51] proposed designing a two-stage R-CNN network with CNN as the main body in the deep learning stage. , Several rectangular boxes were first generated by the candidate box recommendation module. The final face location was then selected with the filtering module. Although Fast R-CNN and Faster R-CNN imply optimized R-CNN algorithms with speedup in the candidate box selection step [52], the aforementioned methods are challenging to accommodate all face detection features. Consequently, Huang et al. [53] recommended the DenseBox algorithm, which uses a full convolutional neural network to simultaneously detect faces and mark face key points. Chen et al. [54], who proposed FacenessNet to train convolutional neural network models for each of the five facial senses, subsequently combined the output information to determine whether the region contains a face.

4.1.2 Face alignment

Further face alignment potentially enhances the performance of facial expression recognition tasks. Asthana proposed using Discriminative Response Map Fitting (DRMF) [55] to boost the face key to 66. SUN et al. [56] recommended the cascaded CNN method to predict key points by cascaded regression apart from considering multiple face detector combinations for high performance. Tasks Constrained Deep Convolutional Network (TCDCN) [57] is a multi-task network for efficient feature point localization, while Multi-task Convolutional Neural Network (MTCNN) [58] recommended a multi-task framework for deep cascading. These methods improve the robustness of face feature point detection through network structure optimization. Recent research has considered integrating distinctive methods. Gyuler [59] recommended combining semantic segmentation with regression networks, while Hu Peiyun [60] integrated shallow and deep features to identify the face location with contextual information. Meanwhile, Wu [61] who drew on human pose estimation presented key points using boundary information.

4.1.3 Data augmentation

Conventional data enhancement methods do not change the original image attributes and model-extracted features and fail to increase the number of feature levels. In this vein, deep learning-oriented data enhancement has garnered much scholarly attention in recent times. Goodfellow proposed Generative Adversarial Network (GAN) [62] to generate high-quality fake images, increase samples, and elicit more diverse features.

Consequently, Mirza et al. [63] proposed that CGAN incorporates category information into the model to generate images with specific labels. Several scholars have also integrated (i) supervised learning CNN with unsupervised learning GAN and (ii) information theory with GAN for optimization purposes [64]. Zhu et al.'s [65] cycleGAN employed unpaired training data to learn a mapping function across domains, which enables images under multiple styles albeit with multiple model training. Choi et al. [66] integrated CGAN with CycleGAN and proposed a StarGAN network to produce multi-style images with a simple generator. Sun et al. [67] later proposed improving the reconstruction error based on StarGAN. The generator employed depth-separable convolution for down-sampling for an enriched variety of expressions.

4.1.4 Face normalization

Related studies have been examined to generate the frontalfacial views of human expressions using pose normalization given the complexities underlying face normalization. For example, Huang et al. [68] recommended a Two-Pathway Generative Adversarial Network (TP-GAN), Wang et al. [69] suggested a Compositional GAN (Comp-GAN), and a GAN-oriented depth model was proposed for face pose transformation with optimal outcomes. Table 3 summarizes the conventional and novel approaches targeting image pre-processing.

Table 3: Comparison between Traditional and New Face Detection Methods

Type		Traditional Method	Deep Learning
Image pre-processing		Edge detection [17] HMM [18]	R-CNN [51] Fast (Faster) R-CNN [52] DenseBox [53] FacenessNet [54]
Face alignment		ASM, AAM [19] SDM [20]	DRMF [55] CDCN [56] TCDCN [57] MTCNN [58] Combination method [59, 60]
Data augmentation		Rotation, tilt, etc. [21] Cropping [22] Noise [23] Distortion [24]	GAN [62] CGAN [63] Combination method [64] CycleGAN [65] StarGAN [66, 67]
Face normalization	Illumination	Histogram [25, 26]	Combination method
	Pose	Global and local perception [28]	TP-GAN [68] Comp-GAN [69]

4.2 Deep Learning Framework

Deep learning-based expression recognition extracts features layer-by-layer through constructive supervised or unsupervised learning. Specifically, supervised learning managed the case of data with labels, which is effective with a broad range of applications. Unsupervised learning manages the case of unlabeled data. Multiple abstract image features are learned and fused into an end-to-end model with higher robustness. Current research emphasizes network architectures entailing feature extraction, which allows the network to train through large datasets and autonomously identify well-learned features. Table 4 demonstrates the excellent deep learning-based FER frameworks developed in recent years.

Table 4: Comparison of Deep Learning Frameworks for Expression Recognition

Items	Deep Belief Network (DBN)	Recurrent Neural Network (RNN)	Convolutional Neural Network (CNN)
Learning Style	Unsupervised	Unsupervised	Supervised
Network Composition	Restricted Boltzmann Machine superposition	Loop and hidden layers	Convolutional, pooling, and fully connected layers
Advantages	Identifying features, classifying data, generating data	Commonly used to process serial data	Parameter reduction by weight sharing, local region linking and down-sampling, direct input to the original image, feature adaptive extraction, rich improvement strategies
Disadvantages	Slow learning process, severe local feature loss during training	Easy gradient disappearance, high computational effort	Requires tuning, large samples, and hardware devices which results in long training time
Application Scenarios	Feature extraction, data dimensionality reduction	Video recognition	Image and video recognition
Improved Model	Convolutional DBN/ Conditional RBM	BRNN/ LSTM	AlexNet/ GoogLeNe/ VGGNet/ ResNet etc.

4.2.1 DBN

Hinton et al. proposed DBN in 2006, which contains multiple neuron layers that are divided into explicit and implicit neurons or elements. Explicit elements receive input, whereas their implicit counterparts extract features following Figure 2 [70]. The model automatically learns features through a multi-layer RBM network, which forms a hierarchical feature model from low-level edge features to abstract high-level subject ones and uses training data to generate the RBM network weight structure or obtain data feature information. Notably, this model is popular in facial expression recognition domains following its ability to automatically learn abstract information from face images.

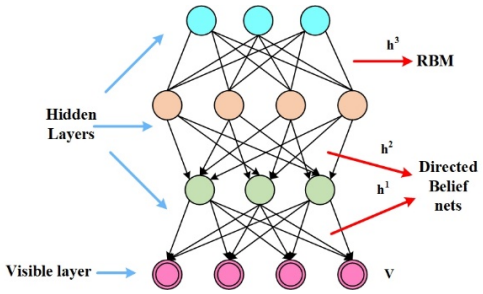

Fig.2 Network Structure of DBN Adapted from [70]

Yang et al. [71] extracted Gabor wavelet features for face samples to be fed into the constructed DBN post-convolutional fusion, while the DBN fine-tuning strategy used the cross-entropy method to determine the optimal model parameters and classified test samples through Softmax regression. The reliability of deep learning methods in face recognition was not demonstrated on large data following insufficient samples and high training time and cost. Huang et al. [72] constructed a new deep confidence network (GB-DBN) with a modified restricted Boltzmann machine (GB-RBM) to learn face expression features, which was used as input to a Stacked Autoencoder (SAE) in recognizing and classifying data samples through seven-fold cross-validation. The GB-DBN+SAE method enhanced the classification robustness albeit with prolonged training time. On another note, Li et al. [73] proposed incorporating CS-LBP with the underlying centrosymmetric principle into DBN while simultaneously regarding the local feature information and texture features of faces through multiple poses. Nevertheless, the overall scheme duration and cost require reduction.

4.2.2 RNN

As a neural network type with a short-term memory capacity, RNN establishes a memory pattern for information storage by adding a loop structure to hidden layer units. The neurons in RNN can receive information from other neurons and amongst themselves to develop a network structure with loops. The model state, which characterizes the acquired historical memory information, is commonly used for face image sequence detection following Figure 3 [74].

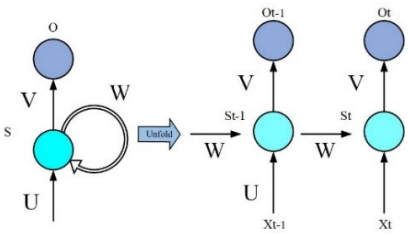

Fig.3 Network Structure of RNN Adapted from [74]

Yang et al. [75] proposed an LSTM- RNN-based dynamic expression recognition method for human faces with an RNN network incorporating LSTM memory cells (cell state) to extract association information from image sequences and make face expression judgments following single-frame images and historical association information. Meanwhile, Zhang et al. [76] suggested a temporal network PHRNN to extract (i) local or whole, (ii) geometric or appearance, and (iii) static or dynamic expression features.

4.2.3 CNN

The CNN, an end-to-end neural network model with features entailing weight-sharing and local connectivity, significantly reduces the network parameters and catalyzes faster training and higher accuracy [77]. Table 5 presents several improved CNN model structures.

Table 5: Comparison between Classic CNN Models

Items	AlexNet	VGGNet	GoogLeNet	ResNet
Time	2012	2014	2014	2015
Layers	8	22	16/19	152
Convolution Kernel Size	11,5,3	3	7,1,3,5	7,1,3,5
Data Enhancement	✓	✓	✓	✓
Dropout	✓	✓	✓	✓
Inception	×	×	✓	×
BN	×	×	×	✓

1) AlexNet

The AlexNet network proposed in 2012, which shifted the research trend from manual features to network architecture, contains four primary parts: five convolutional layers, three pooling layers, two fully connected layers, and one data local normalization layer. Figure 4 below depicts the aforementioned network structure. AlexNet proves advantageous as it utilizes (i) a faster convergence activation function (ReLU), (ii) LRN and dropout to prevent overfitting, and (iii) a maximum pooling method rather than average pooling in the pooling layer. The simplified network structure of AlexNet and application of robust techniques render it valuable for learning and experimentation purposes.

Zhao et al. [78] utilized a feature selection network to filter the features derived from the previous AlexNet network convolutional layer to only retain features that contribute to the emotion. These features were then fed into a Softmax classifier for the expression recognition task, which improved the final expression recognition results while concurrently accelerating the model speed. Yang et al. [79] presented a multi-scale convolution in the AlexNet network, which highly applies to small-sized expression images, extracted feature information at different scales, and fused multiple low-level features with high-level features while passing them downward for an accurate and holistic depiction of the image information.

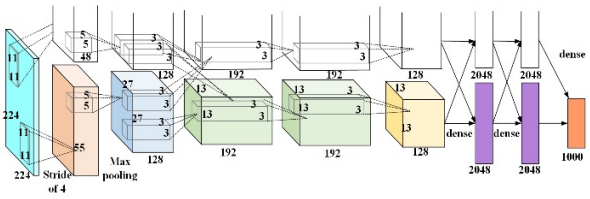

Fig.4 Network Structure of RNN Adapted from [78]

2) VGGNet

The VGGNet network model, which was proposed in 2014, is innovative as the overall network structure employs all small-size convolutional and pooling kernels for optimal performance by increasing the network depth. The increasing number of network layers does not induce a dramatic rise in the number of parameters as neural network parameters are primarily concentrated in the fully connected layers. As such, this structural design substantially reduces the number of parameters while optimizing non-linear operations to improve the network learning ability for image features. Figure 5 illustrates multiple VGGNet model structures: VGG-11, VGG-16, and VGG-19.

Dhankhar et al. [80] combined the ResNet-50 model with that of VGG-16 to structure a novel integrative model to recognize facial expressions and yield optimal outcomes on the KDEF dataset. Meanwhile, Cui et al. [81] proposed an expression recognition algorithm by integrating the improved VGGNet with Focal Loss for an improved Focal Loss. Specifically, a probability threshold was developed to avoid the adverse impact of mislabeled samples on the model classification performance.

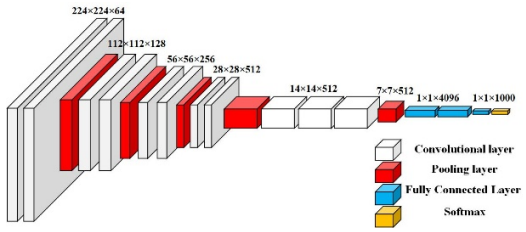

Fig.5 Network Structure of VGG16 Adapted from [80]

3) GoogLeNet

A network structure proposed by the Google team, GoogLeNet, was proposed in the same year as the VGG network to mitigate excessive network parameters by increasing the network structure sparsity. As the Inception structure, the base module is combined in a cascading manner with each module using a different size filter. The module is processed parallel to the maximum pooling operation. Subsequently, the omission of all fully connected layers minimizes the number of parameters and improves the model's computational efficiency and performance. Figure 6 illustrates the module structure. Peng Zhang et al. [82], who yielded positive results by incorporating null convolution into the Inception structure to extract the multi-scale feature information of facial expressions, presented a channel attention mechanism to improve the model's ability and represent important feature information.

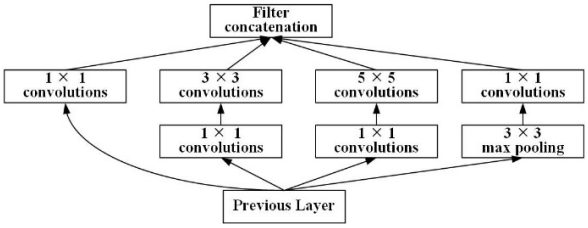

Fig.6 Module Structure of Inception Adapted from [82]

4) ResNet

Hekaiming et al. from Microsoft Research proposed Residual Network (ResNet) in 2015 for accelerated model training, lower error rates, a smaller number of parameters, and little computational effort. Contrary to the traditional convolutional neural network, a residual module is introduced in the network, with a jump connection utilized in the internal residual module. This module effectively alleviates the gradient disappearance problem of back-propagation during network model training to resolve complex deep network training and its poor performance. Figure 7 depicts the basic residual learning unit.

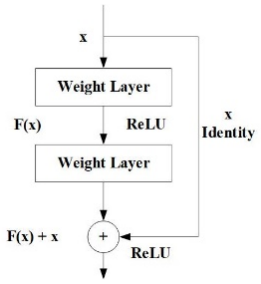

Fig.7 Residual Structure Adapted from [83]

Shen et al. [83] optimized the inverted residual network to be used as a basic unit for lightweight convolutional network model development and integrated the filtered shallow features with deep features for expression recognition. Chen et al. [84] recommended an attention-split convolutional residual network for improved feature representation. The Basic Block in the ResNet18 backbone network was substituted with Coordinate Attention Split Convolution Block (CASCBlock). The number of model parameters was effectively reduced while simultaneously improving the computational speed. Meanwhile, Jiang et al. [85] recommended an expression recognition algorithm with facial key points and weight assignment residual network for the maximum expression range by facial key points to eliminate the interference of image backgrounds and irrelevant content. A weight assignment mechanism was simultaneously introduced to realize the weight assignment of different regions and mitigate inter-class differences.

Other workable models, such as DenseNet, SK-Net, and DANet have emerged from the ILSVRC Challenge and other competitions annually to improvise the basic CNN model to varying degrees. Table 6 presents the performance of representative approaches in deep learning.

Table 6: Performance Summary of Deep Learning-based Representative Methods

Type	Author	Method	Object	Accuracy
DBN	Yang et al. [71]	Gabor +DBN	Small data volume of face images	AR: 92.7%
	Huang et al. [72]	GB-RBM	Distinctive facial images	JAFFE: 92.46%
	Li et al. [73]	CS-LBP+DBN	Distinctive facial images	JAFFE: 92.78%
RNN	Yang et al. [75]	LSTM- RNN	Mixed facial expressions	CK+: 97.7%; MIX: 73.67%
	Zhang et al. [76]	PHRNN	Image sequences	CK+: 97.78%; MMI:

				79.3%
AlexNet	Zhao et al. [78]	FSN+ AlexNet	High-level variation of facial appearance	FER: 67.6%; RAF: 72.4
	Yang et al. [79]	Multi-scale CNN	Small-scale expression images	CK+: 94. 25%; JAFFE: 93.02%
VGGNet	Dhankhar et al.[80]	Multi-Network Convergence	Real environment in the field	KDEF: 80.95%
	Cui et al. [81]	VGGNet +Focal Loss	Misclassification	FER: 72.49%; JAFFE: 97.61%
GoogLeNet	Zhang et al. [82]	Multi-scale Feature Fusion	Suppressing redundant information	FER: 73.32%; CK+: 97.40%
ResNet	Shen et al. [83]	Multi-layer Feature Fusion	Reduction of model parameters	RAF: 85.49%
	Chen et al. [84]	CASCBlock	Enhanced depth of feature extraction	FER: 72.6%; RAF: 84.4%
	Jiang et al. [85]	1.1.1 Weight Allocation	Expressions with minimal interclass variation	FER: 74.14%; CK+: 98.99%

4.3 Deep Learning-Based Expression Recognition

A significant number of face recognition algorithms proposed based on deep learning frameworks could be classified into expression recognition for static images and expression recognition algorithms for dynamic sequences based on the processed data type.

4.3.1 Static images

Static image-based recognition is the mainstream research direction with a large dataset base that is easily used in real-time with single-frame expression images (as input) for expression categories identification through static features. Several researchers emphasized the local regions of facial expressions to capture intricate feature changes and improve facial expression recognition rates. For example, Gao et al. [86] proposed a three-channel TP-FER (tri-path networks for facial expression recognition) method based on optimized convolutional neural networks. Based on the constructed convolutional neural network training, the method utilizes three input channels with an emphasis on the face, eye, and mouth regions for feature extraction and expression discrimination. Specific algorithms also improved the similarity problems in expression recognition by incorporating novel network structures or functions. A lightweight face expression recognition method was proposed by Yin et al. [87] based on convolutional attention. Notably, the decomposed convolution served to reduce the model parameters while embedding the convolutional attention mechanism for facial expression recognition and enhanced model feature extraction capability. Liang et al. [88] presented a compression excitation module in the residual network, assigned weights to the features of different channels, and used different compression rates in different convolutional layers to enhance the network feature extraction ability for facial expressions.

Specific methods were also determined to mitigate the identity change impacts on facial expression recognition by integrating traditional machine learning methods. Ma et al. [89], who proposed a feature fusion algorithm, used the local binary method for the local feature extraction of face expression, global feature extraction of image with the convolutional neural network, and the integration of extracted local expression features with global expression features. Histogram equalization was then employed to process face expression images and perform face expression recognition. Based on the experimental outcomes, this method demonstrated a positive effect on facial expression recognition.

4.3.2 Dynamic Sequence

Expression recognition based on image sequences involves multiple picture frames that convey dynamic shifts of expressions as input and determines expression categories by dynamic features. Some algorithms empirically emphasize micro-expressions through the expression dynamics process. Yu et al. [90] achieved frame aggregation by concatenating the mean, variance, and minimum and maximum values of all frame features. Most sequence-based expression recognition algorithms tend to ignore subtle expressions and expression intensity with training samples of different intensities (as input) and use the intrinsic correlation between expressions in different expression intensity sequences. Thus, optimal feature extraction was performed. Most algorithms prioritize expression processes from subtle to obvious by fusing network architectures or algorithms. Zhang et al. [91] proposed a two-branch network to extract (i) time-domain information from face key points in consecutive frames with a bidirectional recurrent neural network and (ii) space-domain features from face images with a convolutional design warp network. Furthermore, Liu et al. [92] employed

Kim et al.'s [93] recommendation to integrate the network with more intensity states and used five loss functions that minimized the expression classification error, intra-class expression variation, intensity classification error, and intra-intensity variation and coded intermediate intensity for network training regulation. Facial expression analysis through image sequences in actual scenarios is a crucial direction for future research.

5. Facial Expression Recognition Issues, Challenges, and Future Research Direction

5.1 Existing Problems

The issues underpinning facial expression recognition techniques remain relatively unresolved despite much progress in related domains.

1) The reliable expression dataset is small in size, with the expression samples incongruently distributed. For example, the size of the current face expression recognition dataset does not exceed 10,000. Insufficient large-scale and labeled expression data also hamper neural network training with a deep structure for high recognition rates. Multiple factors limit the derivation of large-scale labeled samples, regardless of whether the data are collected in controlled laboratory or uncontrolled natural environments. The images that could be gathered are also very limited given the vast variation in the frequency of different expressions in daily life, thus rendering small-scale training samples and incongruent sample distribution.
2) Face images entail complexities, ambiguities, and other associated factors. The current datasets stem from Europe and America, with few counterparts originating from Asian regions, such as China. The same expression types could also induce vast differences based on how expressions are conveyed between individuals, thus causing large intra-class variations. Additionally, people who occasionally suppress emotional expressions that occur when they are truly emotional could cause ambiguous expressions. The ideal dataset should not only have expression labels, but include other attributes involving age, gender, and race.
3) Regarding face image quality, the pictures acquired in an uncontrolled environment would demonstrate characteristics involving low resolution, high noise, and complex background. The acquired photos also require consideration based on light, noise, pose, and physical occlusion. For example, some of the captured subjects may have beards or wear accessories, such as glasses or hats, thus rendering the captured face images incomplete.
4) The network model is complicated, whereas the training cost is too high. Based on the deep learning model trends characterized by CNN, the number of network layers becomes deeper and deeper for optimal outcomes. Nevertheless, this method of increasing network depth reveals several limitations. Although a large network needs to learn more parameters, which causes network overfitting to the training dataset, adding the number of layers causes an increase in computation and advanced hardware configurations.

5.2 Future Research Direction

1) The demand for real-time recognition and facial expression analysis has become significant with the development of artificial intelligence, computer vision, and other technologies. In this vein, expression recognition has gained a broader research prospect. Following the research status analysis, the development direction of facial expression recognition presents the following development trends.
2) The dataset from laboratory-controlled environments was transferred to real-world and uncontrolled environments. Regardless, actual situations proved more intricate than the ideal experimental data state as the expressions gathered under laboratory conditions are palpable and well-postured with masked faces, uniform lighting, and a single background. As such, real-world-oriented datasets involving field and life scenes are manifested. The challenges underlying expression labeling, expression intricacies, image diversity, and other associated issues require thorough examination following the insufficient data volume.
3) The algorithmic framework of facial expression recognition based on deep learning was studied to combine the latest artificial intelligence theories. As the features of facial expression images demonstrate different application environments with specific benefits and drawbacks, the use of a single method may instigate information loss. It is deemed necessary to aggregate different techniques and enhance the model robustness to effectively improve the expression recognition rate.
4) The facial expression recognition models were examined based on cross-disciplinary knowledge. Face recognition development is promoted by integrating advanced techniques from biology, psychology, statistics, and other disciplines. For example, the eye-movement technology and attention mechanism proposed in clinical medicine and human brain vision, respectively, effectively exemplify cross-border collaboration.

REFERENCES

[1] Mehrabian. A, "*Nonverbal communication,*" Routledge, 2017, pp.193-200.

[2] Ekman. P and Heider. K.G, "*The universality of a contempt expression: A replication,*" Motiv Emot 12, 1988, pp.303–308.

[3] S. Li and W. Deng, "*Deep Facial Expression Recognition: A Survey,*" in IEEE Transactions on Affective Computing, vol.13, no.3, 2022, pp.1195-1215.

[4] M. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba, "*Coding facial expressions with Gabor wavelets,*" Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 200-205.

[5] T. Kanade, J. F. Cohn and Yingli Tian, "*Comprehensive database for facial expression analysis,*" Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), 2000, pp. 46-53, doi: 10.1109/AFGR.2000.840611.

[6] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "*The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression,*" 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 94-101, doi: 10.1109/CVPRW.2010.5543262.

[7] M. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba, "*Coding facial expressions with Gabor wavelets,*" Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 200-205, doi: 10.1109/AFGR.1998.670949.

[8] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang and M. J. Rosato, "*A 3D facial expression database for facial behavior research,*" 7th International Conference on Automatic Face and Gesture Recognition (FGR06), 2006, pp. 211-216, doi: 10.1109/FGR.2006.6.

[9] Moreno. A and Sanchez. A, "*Gavabdb: A 3d face database,*" In 2nd COST275 workshop on Biometrics on the Internet, 2004, pp. 75-80.

[10] X. Li, T. Pfister, X. Huang, G. Zhao and M. Pietikäinen, "*A Spontaneous Micro-expression Database: Inducement, collection and baseline,*" 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013, pp.1-6, doi: 10.1109/FG.2013.6553717.

[11] Yan W-J, Li X, Wang S-J, Zhao G, Liu Y-J, Chen Y-H, et al., "*CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation,*" PLoS, vol.9, no.1, 2014.

[12] Goodfellow, I.J. et al., "*Challenges in Representation Learning: A Report on Three Machine Learning Contests,*" In Neural Information Processing(ICONIP), vol.8228, 2013.

[13] Shan Li, Weihong Deng, and JunPing Du, "*Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,*" In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2852-2861.

[14] Dhall, R. Goecke, S. Lucey, and T. Gedeon, "*Static facial expressionanalysis in tough conditions: Data, evaluation protocol and benchmar*k," in Computer Vision Workshops (ICCV Workshops), 2011 IEEE Interna-tional Conference on. IEEE, 2011, pp. 2106–2112.

[15] Dhall, R. Goecke, S. Lucey, T. Gedeon et al., "*Collecting large, richlyannotated facial-expression databases from movie*s," IEEE multimedia, vol. 19, no. 3, 2012, pp. 34–41.

[16] Hallinan PW, Gordon G, Yuille AL, Giblin P, and Mumford D, "*Two- and Three-Dimensional Patterns of the Face (1st ed.),*" CRC Press, 1999.

[17] S. Phimoltares, C. Lursinsap and K. Chamnongthai, "*Face detection and facial feature localization without considering the appearance of image context*," Image and Vision Computing,Vol. 25, no. 5,2007,pp.741-753, doi:10.1016/j.imavis.2006.05.017.

[18] Wang, Jilin, Ye, Jian-Long, Zhao, Li, and Zou, Cai-Rong, "*Face Detection Based on Curvilinear Wave Hidden Markov Model,*" Journal of Sensing Technology, vol. 24, no.5, 2011, pp.714-718.

[19] T.F. Cootes, C.J. Taylor, D.H. Cooper, and Graham. J, "*Active Shape Models-Their Training and Application,*" Computer Vision and Image Understanding, Vol. 61, no.1, 1995, pp.38-59.

[20] Xiong. X and De la Torre. F, "*Supervised descent method and its applications to face alignment,*" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 532-539

[21] Simard, Patrice Y., David Steinkraus, and John C. Platt, "*Best practices for convolutional neural networks applied to visual document analysis,*" Icdar, vol.3, 2003.

[22] Krizhevsky. A, Sutskever. I, and Hinton. G. E, "*Imagenet classification with deep convolutional neural networks,*" Communications of the ACM, vol.60, no.6, 2017, pp.84-90.

[23] André Teixeira Lopes, Edilson de Aguiar, Alberto F. De Souza, and Thiago Oliveira-Santos, "*Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order,*" Pattern Recognition, vol. 61, 2017, pp. 610-628.

[24] M.V. Zavarez, R.F. Berriel and T. Oliveira-Santos, "*Cross-Database Facial Expression Recognition Based on Fine-Tuned Deep Convolutional Network,*" 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2017, pp. 405-412.

[25] Yu. Z, and Zhang. C, "*Image based static facial expression recognition with multiple deep network learning,* " In Proceedings of the 2015 ACM on international conference on multimodal interaction,2015, pp. 435-442.

[26] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee, "*Hierarchical committeeof deep cnns with exponentially-weighted decision fusion for staticfacial expression recognition,*" in Proceedings of the 2015 ACM onInternational Conference on Multimodal Interaction. ACM, 2015, pp. 427–434.

[27] J. Li and E. Y. Lam, "*Facial expression recognition using deep neural networks,*" 2015 IEEE International Conference on Imaging Systems and Techniques (IST), 2015, pp. 1-6, doi: 10.1109/IST.2015.7294547.

[28] Rui Huang, Shu Zhang, Tianyu Li, and Ran He, "*Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis,*" in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2439-2448

[29] M. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba, "*Coding facial expressions with Gabor wavelets,*" in Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 200-205, doi: 10.1109/AFGR.1998.670949.

[30] Zhen Wen and Huang, "*Capturing subtle facial motions in 3D face tracking,*" in Proceedings Ninth IEEE International Conference on Computer Vision, vol.2, 2003, pp.1343-1350, doi: 10.1109/ICCV.2003.1238646.

[31] Feng. X, Pietikäinen. M, and Hadid. A, "*Facial expression recognition based on local binary patterns,*" Pattern Recognit Image Anal, vol. 17, 2007, pp. 592–598.

[32] Yimo Guo, Guoying Zhao, and Matti Pietikäinen, "*Discriminative features for texture description,*" Pattern Recognition, vol.45, no.10, 2012, pp.3834-3843, doi:10.1016/j.patcog.2012.04.003.

[33] Satiyan. M, Hariharan. M, and Nagarajan. R, "*Recognition of facial expression using Haar wavelet transform,*" Journal of Electrical and Electronics Systems Research (JEESR), Vol.3, no.11, 2010, pp. 89-96.

[34] M. Hu, K. Li, XH. Wang, and FJ. Ren, "*Facial expression recognition based on histogram weighted HCBP,*" Journal of Electronic Measurement and Instrument, vol.29, no.7, 2015, pp. 953-960.

[35] M. Pantic and L. J. M. Rothkrantz, "*Facial action recognition for facial expression analysis from static face images,*" in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 34, no.3, 2004, pp. 1449-1461, doi: 10.1109/TSMCB.2004.825931.

[36] David Cristinacce, Tim Cootes, and Ian Scott, "*A multi-stage ap-proach to facial featuredetection,*" Proceedings of the British Machine Vision Conference, 2004, pp.231-240.

[37] Y. Saatci and C. Town, "*Cascaded classification of gender and facial expression using active appearance models,*" in 7th International Conference on Automatic Face and Gesture Recognition (FGR06), 2006, pp. 393-398, doi: 10.1109/FGR.2006.29.

[38] L. Ma and K. Khorasani, "*Facial expression recognition using constructive feedforward neural networks,*" in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 34, no. 3, pp. 1588-1595, June 2004, doi: 10.1109/TSMCB.2004.825930.

[39] Wu. XG, and Luo. LM, "*An improved optical flow field calculation method,*" Acta electronica sinica, Vol. 1, 2000, pp.130-131.

[40] LI Dan, XIE Lun, LU Ting, HAN Jing, HU Bo, WANG Zhi-liang, and REN Fu-ji, "*Capture of micro-expressions based on the entropy of oriented optical flow,*" Chinese Journal of Engineering, vol. 39, no. 11, 2017, pp.1727-1734.

[41] Huang YM, Zhang GB, Dong F, and Da FP, "*Gabor-based, Fisher face multi-feature extraction and integrated SVM for face expression recognition,*" Computer Application Research, vol. 28, no. 4, 2011, pp.1536-1539.

[42] Liu. P, Zhou. J.T, Tsang. I.WH, Meng. Z, Han. S, and Tong. Y, "*Feature Disentangling Machine - A Novel Approach of Feature Selection and Disentangling in Facial Expression Analysis,*" Lecture Notes in Computer Science, vol. 8692, 2014, pp.151-166.

[43] Qi et al., "*Facial Expressions Recognition Based on Cognition and Mapped Binary Patterns,*" in IEEE Access, vol. 6, pp. 18795-18803, 2018, doi: 10.1109/ACCESS.2018.2816044.

[44] Xu WH and Sun ZX, "*An LSVM algorithm for video sequence expression classification,*" Journal of Computer-Aided Design and Graphics, vol. 21, no. 4, 2009, pp.542-548.

[45] WANG Xiaohua, XIA Chen, and Hu Min, "*Video sequence facial expression recognition with spatio-temporal features,*" Journal of electronics and information, vol. 40, no.3, 2018, pp.626-632.

[46] K R. Kandemir and G. Özmen, "*FACIAL EXPRESSION CLASSIFICATION WITH HAAR FEATURES, GEOMETRIC FEATURES AND CUBIC BÃ‰ZIER CURVES*", IU-Journal of Electrical & Electronics Engineering, vol. 13, no. 2, 2013, pp.1667-1673.

[47] W. Lirong, Y. Xiaoguang, W. Jianlei, X. Jing and Z. Jian, "*Facial Expression Recognition Based on Local Texture Features,*" in 2011 14th IEEE International Conference on Computational Science and Engineering, 2011, pp. 543-546, doi: 10.1109/CSE.2011.96.

[48] Zhan. YZ, Cheng. KY, Chen. YB, et al., "*A New Classifier for Facial Expression Recognition: Fuzzy Buried Markov Model,*" in J. Comput. Sci. Technol. vol. 25, no. 3, 2010, pp. 641–650.

[49] N. Devi and M. V. V. R. M. K. Rao, "*A novel method to achieve optimization in facial expression recognition using HMM,*" in 2015 International Conference on Signal Processing and Communication Engineering Systems, 2015, pp. 48-52.

[50] J. See, "*Probabilistic Bayesian network classifier for face recognition in video sequences,*" in 2011 11th International Conference on Intelligent Systems Design and Applications, 2011, pp. 888-893.

[51] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "*Rich feature hierarchies for accurate object detection and semantic segmentation,*" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580-587.

[52] Ren. S, He. K, Girshick. R, and Sun. J, "*Faster R-CNN: Towards real-time object detection with region proposal networks,*" Advances in neural information processing systems, Vol. 28, 2015.

[53] Huang. L, Yang. Y, Deng. Y, and Yu. Y, "*Densebox: Unifying landmark localization with end to end object detection,*" arXiv preprint arXiv:1509.04874, 2015.

[54] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang, "*From facial parts responses to face detection: A deep learning approach,*" in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3676-3684.

[55] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic, "*Robust discriminative response map fitting with constrained local models,*" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3444-3451.

[56] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "*Deep convolutional network cascade for facial point detection,*" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3476-3483.

[57] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "*Facial landmark detection by deep multi-task learning,*" In European conference on computer vision(ECCV), 2014, pp.94-108.

[58] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "*Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks,*" in IEEE Signal Processing Letters, vol.23, no.10, 2016, pp. 1499-1503, doi: 10.1109/LSP.2016.2603342.

[59] Riza Alp Guler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos, "*DenseReg: Fully Convolutional Dense Shape Regression In-The-Wild,*" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6799-6808

[60] Hu P, and Ramanan D, "*Finding tiny faces,*" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp.951-959.

[61] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou, "*Look at Boundary: A Boundary-Aware Face Alignment Algorithm,*" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2129-2138

[62] Goodfellow. I, Pouget-Abadie. J, Mirza. M, Xu. B, Warde-Farley. D, Ozair. S, ... and Bengio. Y, "*Generative adversarial networks,*" Communications of the ACM, vol. 63, no. 11, 2020, PP.139-144.

[63] Mirza M, and Osindero S, "*Conditional Generative Adversarial Nets,*" arXiv e-prints, 2014, arXiv: 1411.1784.

[64] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel, "*InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets,*" Advances in Neural Information Processing Systems 29 (NIPS),Vol 29,2016.

[65] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "*Unpaired image-to-image translation using cycle-consistent adversarial networks,*" in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2223-2232.

[66] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo, "*StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation,*" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8789-8797.

[67] SUN X, and DING X, "*Data augmentation method based on generative adversarial networks for facial expression recognition sets,*" Computer Engineering and Applications, vol. 56, no. 4, 2020, pp.115-121.

[68] Rui Huang, Shu Zhang, Tianyu Li, and Ran He, "*Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis,*" in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2439-2448.

[69] Wang W, Sun Q, Fu Y, Chen T, Cao C, Zheng Z, ... ,and Xue X, "*Comp-GAN: Compositional generative adversarial network in synthesizing and recognizing facial expression,*" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8789-8797

[70] Susskind. JM, Hinton. GE, and Movellan. JR, "*Generating facial expressions with deep belief nets,*" Affective Computing. InTech, 2008, pp.421–440.

[71] Yang R, Zhang YW, and Gou S, "*Gabor Features Combined with Deep Belief Networks for Face Recognition,*" Sensors and Microsystems, vol. 36, no. 5, 2017, pp.68-70.

[72] Huang Shouxi, and Qiu Weigen, "*Face expression recognition based on improved deep belief network,*" Computer Engineering and Design, vol. 38, no. 6, 2017, pp.1580-1584.

[73] Li. C, Wei. W, Wang. J, Tang. W, and Zhao. S, "*Face Recognition Based on Deep Belief Network Combined with Center-Symmetric Local Binary Pattern,*" Advanced Multimedia and Ubiquitous Engineering, vol 393, 2016, pp.277-283.

[74] G. Williams, R. Baxter, Hongxing He, S. Hawkins and Lifang Gu, "*A comparative study of RNN for outlier detection in data mining,*" 2002 IEEE International Conference on Data Mining, 2002, pp. 709-712, doi: 10.1109/ICDM.2002.1184035.

[75] Yang Y, Fang D, and Zhu D, "*Facial expression recognition using deep belief network,*" Rev. Tec. Ing. Univ. Zulia, vol.39, no.2, 2016, pp.384-392.

[76] K. Zhang, Y. Huang, Y. Du and L. Wang, "*Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks,*" in IEEE Transactions on Image Processing, vol. 26, no. 9, 2017, pp. 4193-4203, doi: 10.1109/TIP.2017.2689999.

[77] Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, "*Fuzzy commonsense reasoning for multimodal sentiment analysis,*" PatternRecognition Letters, vol.125, 2019, pp.264-270.

[78] S Zhao, H Cai, H Liu, J Zhang, and S Chen, "*Feature Selection Mechanism in CNNs for Facial Expression Recognition,*" in BMVC, 2018, pp. 317.

[79] Yang X and Shang ZH, "*Facial expression recognition based on improved AlexNet,*" Laser & Optoelectronics Progress, vol.57, no.14, 2020, pp.243-250.

[80] Dhankhar. P, "*ResNet-50 and VGG-16 for recognizing facial emotions,*" International Journal of Innovations in Engineering and Technology (IJIET), vol.13, no.4, 2019, pp.126-130.

[81] Cui Ziyue, Pi Jiatian, Chen Yong, Yang Jiezhi, Xian Yan and Wu Zhiyou, "*Combining improved VGGNet and focal loss for facial expression recognition,*" Computer Engineering and Applications, vol.57, no.19, 2021, pp.171-178.

[82] SHI Hao, XING Yuhang, and CHEN Lian, "*Facial expression recognition based on multi-scale feature fusion and attention mechanism,*" Microelectronics & Computer, vol.39, no.3, 2022, pp.34-40.

[83] Shen Hao, Meng Qinghao, and Liu Yinbo, "*Face expression recognition based on multilayer feature fusion with lightweight convolutional networks,*" Advances in Lasers and Optoelectronics, vol.58, no.6, 2021.

[84] Chen Jiamin and Xu Yang, "*Expression recognition based on attention splitting convolution residual network,*" Laser & Optoelectronics Progress, vol. 59, no.18, 2022.

[85] JIANG Yuewu, ZHANG Yujin, and SHI Jianxin, "*Expression Recognition Method Combining Key Points and Residual Network of Weight Allocation,*" Computer Engineering and Applications, vol.58, no.17, 2022, pp.181-188.

[86] Gao Jingwen, Cai Yongxiang, and He Zongyi, "*TP-FER: A Three-Channel Face Expression Recognition Method Based on Optimized Convolutional Neural Network,*" Computer Applications Research, 2021.

[87] Pengbo Yin, Weimin Pan, and Navy Zhang, "*A lightweight method for face expression recognition based on convolutional attention,*" Advances in Lasers and Optoelectronics, vol.58, no.12, 2021.

[88] Liang Huagang and Lei Yixiong, "*Enhanced Separable Convolutional Channel Features for Expression Recognition,*" Computer Engineering and Applications, vol. 58, no.2, 2022, pp.184-192.

[89] Li Zhe, Li Keqing, and Deng Junyong, "*Research on face expression recognition method based on multi-feature combination,*" Computer Application Research, vol.4, 2020.

[90] T. Yu, J. Yu, Z. Yu and D. Tao, "*Compositional Attention Networks With Two-Stream Fusion for Video Question Answering,*" in IEEE Transactions on Image Processing, vol. 29, 2020, pp. 1204-1218, doi: 10.1109/TIP.2019.2940677.

[91] K. Zhang, Y. Huang, Y. Du and L. Wang, "*Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks,*" in IEEE Transactions on Image Processing, vol. 26, no. 9, 2017, pp.4193-4203.

[92] Liu C, Tang T, Lv K, and Wang M, "*Multi-feature based emotion recognition for video clips,*" in Proceedings of the 20th ACM International Conference on Multimodal Interaction, 2018, pp. 630-634.

[93] D. H. Kim, W. J. Baddar, J. Jang and Y. M. Ro, "*Multi-Objective Based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition,*" in IEEE Transactions on Affective Computing, vol.10, no.2, 2019, pp. 223-236, doi: 10.1109/TAFFC.2017.2695999.

YAN BOWEN received his B.S. in Electronic Information Engineering in 2014 and M.S. in Electronics and Communication Engineering in 2017. He is currently pursuing his Ph.D. degree in Computer Science at Taylor's University with image processing and deep learning as research interests.

Azween Abdullah is a professional development alumnus of Stanford University and MIT with 30 years of experience as an academic in institutions of higher learning and 15 years as a software engineer, systems analyst, computer software developer, and IT/MIS consultant and trainer in commercial companies. He was also the director of research and academic affairs at two institutions of higher learning and vice president of educational consultancy services.

Lorita Angeline has published research on computer vision, intelligent transportation system (ITS), artificial intelligence (AI), and pattern recognition projects. Dr. Angeline, who has 14 years of teaching experience, is currently a senior lecturer at Taylor's University.

SH Kok has 17 years of industry experience before completing his Ph.D. in Computer Science from Taylor's University in 2021 and currently works as a senior lecturer.