symmetry

MDPI

Article

Facial Expression Recognition: A Survey

Yunxin Huang, Fei Chen, Shaohe Lv and Xiaodong Wang *

College of Computer, National University of Defense Technology, Changsha 410073, China;
hallie9499@126.com (Y.H.); chenfei14@nudt.edu.cn (F.C.); shaohelv@nudt.edu.cn (S.L.)
* Correspondence: xdwang@nudt.edu.cn

Received: 2 September 2019; Accepted: 16 September 2019; Published: 20 September 2019

Abstract: Facial Expression Recognition (FER), as the primary processing method for non-verbal intentions, is an important and promising field of computer vision and artificial intelligence, and one of the subject areas of symmetry. This survey is a comprehensive and structured overview of recent advances in FER. We first categorise the existing FER methods into two main groups, i.e., conventional approaches and deep learning-based approaches. Methodologically, to highlight the differences and similarities, we propose a general framework of a conventional FER approach and review the possible technologies that can be employed in each component. As for deep learning-based methods, four kinds of neural network-based state-of-the-art FER approaches are presented and analysed. Besides, we introduce seventeen commonly used FER datasets and summarise four FER-related elements of datasets that may influence the choosing and processing of FER approaches. Evaluation methods and metrics are given in the later part to show how to assess FER algorithms, along with subsequent performance comparisons of different FER approaches on the benchmark datasets. At the end of the survey, we present some challenges and opportunities that need to be addressed in future.

Keywords: facial expression recognition; feature extraction; classification; deep learning

1. Introduction

Facial expression is a major non-verbal means of expecting intentions in human communication. The work of Mehrabian [1] in 1974 shows that 55% of messages pertaining to feelings and attitudes is in facial expression, 7% of which is in the words that are spoken, the rest of which are paralinguistic (the way that the words are said). Facial expression has proven to play a vital role in the entire information exchange process in Mehrabian's findings. With the rapid development of artificial intelligence, automatic recognition of facial expressions has been intensively studied in recent years. The study of Facial Expression Recognition (FER) has received extensive attention in the fields of psychology, computer vision, and pattern recognition. FER has broad applications in multiple domains, including human–computer interaction [2,3], virtual reality [4], augmented reality [5], advanced driver assistance systems [6,7], education [8], and entertainment [9].

Various kinds of data can be used as the input of the emotion recognition system. Expression recognition and emotion recognition is related but different. Human facial image is the mainstream and promising input type, because it can provide abundant information for expression recognition research. Besides the facial image taken by camera, physiological signals [10], e.g., electromyograph (EMG), electrocardiogram (ECG), electroencephalograph (EEG), can also be employed as the auxiliary data source in some real-world FER applications.

1.1. Research Background of FER

In the decades of studying FER, Action Units (AUs) [11] and the Valence–Arousal space (V–A space) [12] are two popular models. The V–A space is a universal model widely used in continuous emotion recognition tasks of audio, visual, and physiological signals. As shown in Figure 1, the V–A model identify emotion categories according to the value of the emotion dimensions (i.e., arousal and valence). The AUs encode basic movements of facial muscles, and the combination of AUs could be utilised for FER. In [13], a framework is proposed to apply AUs to estimate the V–A intensity.

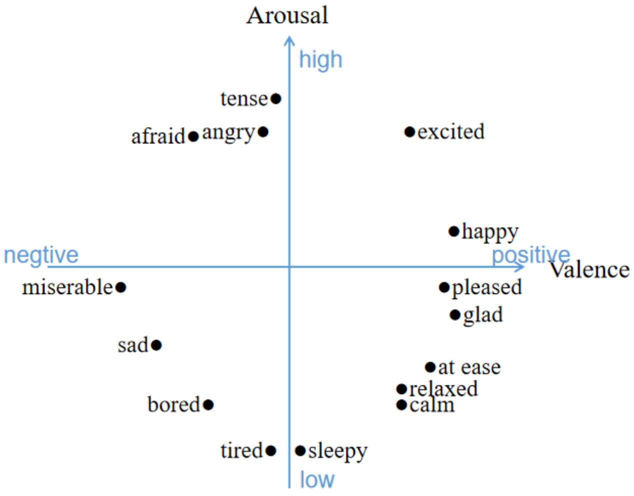


Figure 1. Various emotions and Valence-Arousal Space.

Considering that this survey only focuses on FER of visible facial expressions, the following introduction and discussion are based on the AUs. Studies can be divided into two groups according to whether the features are manually extracted or generated through the output of neural networks, i.e., the conventional FER approaches and the deep learning-based FER approaches.

The conventional FER approach is composed of three major steps, i.e., image preprocessing, feature extraction, and expression classification. Such methods based on manual feature extraction are less dependent on data and hardware, which have advantages in small data sample analysis.

Deep learning-based FER approaches greatly reduce the dependence on feature extraction by employing an "end-to-end" learning directly from input data to classification result. Note that, massive datasets with annotations are the cornerstone of a deep learning algorithm, otherwise, overfitting can easily occur.

According to the shooting environment, FER-related data can be approximately divided into laboratory type and wild type, and several publicly available FER datasets are described in detail in Section 4. Most of the existing studies are based on laboratory datasets, such as JAFFE [14] and CK+ [15], which data come from volunteers who make corresponding expressions under particular instructions. Experiments based on these kind of data accelerate the advancement of FER algorithms. However, with the development of artificial intelligence technology and the wide demand for applications in the era of big data, studies on FER will focus on the spontaneous expressions in the wild. New solutions to FER in complex environment, e.g., occlusion, multi-view, and multi-objective, are necessary to be proposed.

1.2. Terminologies

Before reviewing the approaches of FER, we first present some related terminologies to supplement the theoretical basis of FER technology. Facial Landmarks (FLs), Facial Action Units (AUs), and Facial Action Coding System (FACS) are about how to convert facial action into expression.

Basic Emotions (BEs), Compound Emotions (CEs), and Micro Expressions (MEs) are different definition criteria for expression categories. Existing studies on FER are based on the setting of these concepts and terms.

1.2.1. Facial Landmarks (FLs)

Facial landmarks [16,17] are visually highlights in facial area, such as the alae of the nose, the end of the eyebrow, and the mouth corner, as shown in Figure 2. The locations of the FLs around facial components and contour, capturing facial deformations due to head movements and facial expressions. Point-to-point correspondences of facial landmarks can establish a feature vector of a human face.

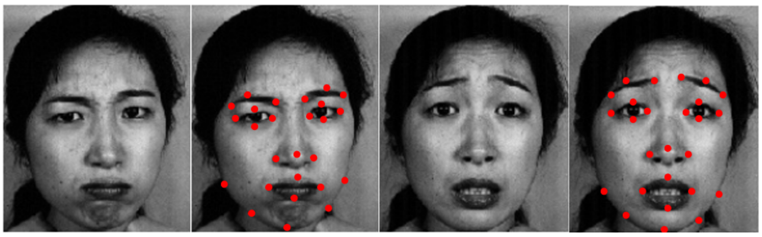


Figure 2. Example of facial landmarks (face images are taken from JAFFE dataset [14]).

1.2.2. Facial Action Units (AUs)

The 46 facial action units encode the basic movements of individual or groups of muscles that are typically observed when a facial expression produces a particular emotion [11]. Figure 3 illustrates some examples. The FER system classifies expression categories by inspecting the combinations of the detected face AUs. For instance, if an image is annotated with 1, 2, 5, and 25 AUs, it can be classified as an exmotion that expresses the "Awed" category.

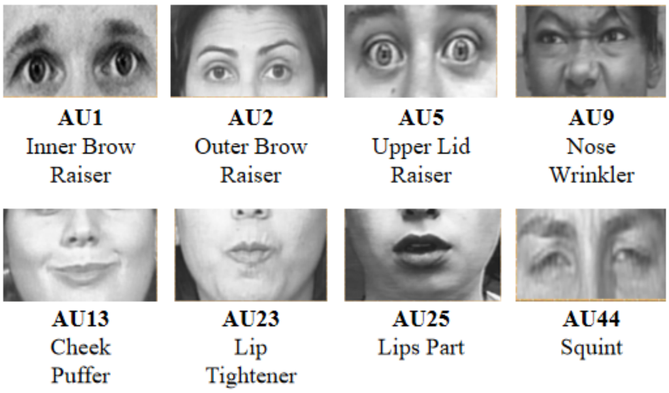


Figure 3. Some examples of AUs (images are taken from CK+ dataset [15]).

1.2.3. Facial Action Coding System (FACS)

Internationally renowned psychologists Ekman and Friesen portray the correspondence between facial muscle movements and expressions through observations and biofeedback [18]. Based on anatomical features, they first divide the whole face into several independent and interrelated AUs, and further analyse the characteristics of these AUs. The typical AUs seen in each of the basic and compound emotion categories are shown in Table 1. The FACS system classifies many human expressions in real life, and is the definitive reference standard for muscle movements in facial expressions today.

Table 1. Prototypical AUs seen in basic and compound emotion category, proposed in [19].

Category	AUs	Category	AUs
Happy	12,25	Sadly disgusted	4,10
Sad	4,15	Fearfully angry	4,20,25
Fearful	1,4,20,25	Fearfully surprised	1,2,5,20,25
Angry	4,7,24	Fearfully disgusted	1,4,10,20,25
Surprised	1,2,25,26	Angrily disgusted	4,25,26
Disgusted	9,10,17	Disgusted surprised	1,2,5,10
Happily sad	4,6,12,25	Happily fearfully	1,2,12,25,26
Happily surprised	1,2,12,25	Angrily disgusted	4,10,17
Happily disgusted	10,12,25	Awed	1,2,5,25
Sadly fearful	1,4,15,25	Appalled	4,9,10
Sadly angry	4,7,15	Hatred	4,7,10
Sadly surprised	1,4,25,26		

1.2.4. Basic Emotions (BEs)

Six basic human emotions, i.e., happiness, surprise, sadness, anger, disgust, and fear, are proposed in [20]. FER-related datasets are generally labelled with these six BEs.

1.2.5. Compound Emotions (CEs)

Compound emotions are combinations of two basic emotions. Twenty two emotions are introduced in [21], including 7 basic emotions (6 basic emotions and 1 neutral), 12 compound emotions expressed commonly by humans, and 3 additional emotions (Awed, Appalled, and Hatred).

1.2.6. Micro Expressions (MEs)

Micro expressions [22] represent more spontaneous and subtle facial movements that occur involuntarily. They tend to reveal the true and potential expressions of a person for a limited time. The duration of the micro expression is very short and lasts for only 1/25 to 1/3 s. Studies on micro expressions are often applied in psychology and police investigations.

1.3. Differences with Existing Survey and Contributions

Currently, there are some other surveys which summarise the development of FER techniques. However, those studies either focus on traditional FER methods or are limited to one specific aspect of FER. In [23], Samal et al. first work to survey the studies of recognition and analysis of face and facial expressions, and the analysis of facial expressions is only a small part of their work. Fasel et al. subsequently focus on automatic facial expression in [24], in which different stages of facial expression analysis system are analysed separately, and some representative FER systems are introduced. The work in [25,26] concentrate on 3D FER. They present currently available 3D/4D face databases and the existing 3D/4D data-oriented FER systems in detail. [27] targets facial micro-expression recognition. Geometry-based and a appearance-based automatic FER systems are utilised in [28]. The work in [29] is about face detection and recognition, and facial expression analysis based on Local Binary Patterns (LBP). Zhang et al. [30] concentrate on the problem of partial occlusion in FER and outline existing challenges and possible opportunities. Unlike most studies based on existing images or videos, Deshmukh et al. [31] summarise the latest advances in the algorithms and techniques used in distinct phases of real-time FER. Some real-life challenges of FER systems are proposed in [32,33].

This article is a comprehensive survey of FER. We propose the general framework for conventional FER approaches and survey new state-of-the-art deep learning-based FER approaches. In addition, to the best of our knowledge, this is the first work to broadly review the FER datasets and summarise the elements of datasets that are related to the choosing and processing of FER methods. Moreover, some challenges and opportunities for future research are further proposed.

1.4. Organisation of the Survey

The rest of this article is organised as follows. Conventional FER approaches are described in terms of image preprocessing, feature extraction, and expression classification in Section 2. Advanced deep-learning-based FER approaches are introduced in Section 3. Several publicly available FER datasets with discussion are provided in Section 4. An overview of the evaluation metrics and performance comparisons are presented in Section 5. Moreover, some challenges and opportunities are mentioned and discussed in Section 6. Finally, we conclude this article in Section 7.

2. Conventional FER Approaches

A notable characteristic of the conventional FER approach is that it is highly dependent on manual feature engineering. The researchers need to preprocess the image and select the appropriate feature extraction and classification method for the target dataset. The conventional FER procedure can be divided into three major steps: image preprocessing, feature extraction, and expression classification, as shown in Figure 4.

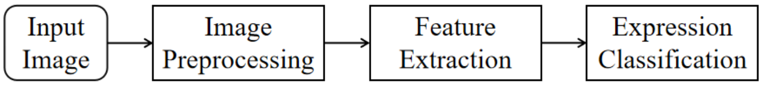


Figure 4. Procedure in conventional FER approaches.

2.1. Image Preprocessing

This step is to eliminate irrelevant information of input images and enhance the detection ability of relevant information. Image preprocessing can directly affect the extraction of features and the performance of expression classification. For various reasons, pictures are often contaminated by some other signals. Some pictures may still have complex backgrounds, e.g., light intensity, occlusion, and other interference factors, even if they are basically free of noise. Moreover, many datasets are different in size, and some are composed of colour images, while some are composed of grayscale images. In addition, various shooting equipment can cause data diversity. These objective interference factors need to be preprocessed before recognition.

The process of image preprocessing are introduced as follows.

- **Noise reduction** is the first preprocessing step. Average Filter (AF), Gaussian Filter (GF), Median Filter (MF), Adaptive Median Filter (AMF), and Bilateral Filter (BF) are frequently used image processing filters.
- **Face detection** has developed into an independent field [34,35]. It is an essential pre-step in FER systems, with the purpose of localising and extracting the face region.
- **Normalisation** of the scale and grayscale is to normalise size and colour of input images, the purpose of which is to reduce calculation complexity under the premise of ensuring the key features of the face [36–38].
- **Histogram equalisation** is applied to enhance the image effect [39].

2.2. Feature Extraction

Feature extraction is a process to extract useful data or information from the image, e.g., values, vectors, and symbols. These extracted "non-image" representations or descriptions are features of the image. Widely used feature extraction methods in FER systems mainly include Gabor feature extraction, Local Binary Pattern (LBP), optical flow method, Haar-like feature extraction, feature point tracking, etc.

Feature extraction may directly influence the performance of the algorithms, which is usually the bottleneck of the FER system. It is essential to take both applicability and feasibility into consideration when manually choosing an appropriate feature extraction method in conventional FER approaches.

2.2.1. Gabor Feature Extraction

The Fourier Transform-based Gabor wavelet kernel function is proposed by combining the wavelet theory with the Gabor feature. Combined with other classification methods [40], FER based on Gabor wavelets has significant advantages. In [14], a set of Gabor filters, which are multi-orientation and multi-resolution, are used to code facial expression images. Yu et al. [41] use linear and nonlinear synthesis of new algorithms on the basis of Gabor feature. Gabor-mean-DWT (Discrete Wavelet Transform) [42] provides a more compact feature vector compared to existing Gabor-based expression classification to alleviate the problem of dimensionality.

Gabor wavelet has good robustness to multi-scale and multi-directional texture feature transformation, and is not sensitive to illumination intensity. Nevertheless, Gabor wavelets may consume a lot of memory because it usually works on global features.

2.2.2. Local Binary Pattern (LBP)

The LBP [43] calculates the brightness relationship between each pixel contained in the image and its local neighbourhood. The binary sequence is then encoded to form a local binary pattern. Finally, it uses a multi-region histogram as a feature description of the image, as shown in Figure 5.

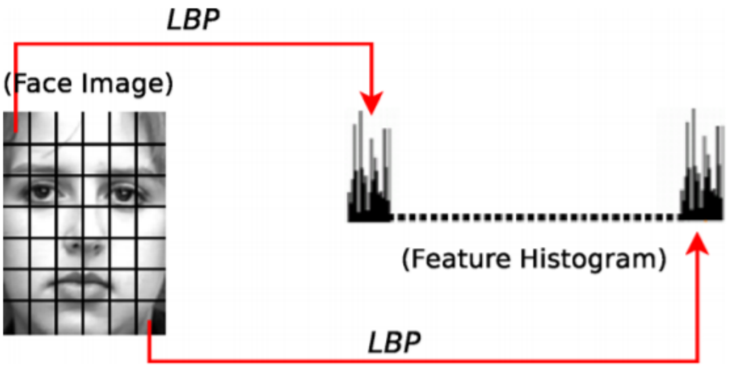


Figure 5. Extraction of LBP histogram from a facial image.

Feng et al. [44] establish local LBP histograms and connect them to FER. An improved LBP-based algorithm, i.e., Complete Local Binary Pattern (CLBP), is proposed in [45], which has better performance than original LBP algorithm, but it also causes dimensionality disaster. Jabid and Chae [46] introduce LBP-based Local Directional Pattern (LDP) algorithm, which is robust to illumination and has relatively low computational complexity. In addition, Local Phase Quantisation (LPQ) [47] is mainly based on short-time Fourier Transform and is stable in feature extraction. In [48], the improved es-LBP (expression-specific LBP) feature is proposed to extract the spatial information, the cr-LPP (class-regularised Locality Preserving Projection) method is proposed to simultaneously maximise the class independence and preserve the local feature similarity.

Compared with Gabor wavelet, the LBP operator requires less storage space and has higher computational efficiency. However, the LBP operator is ineffective on the images with noise. It may lose some useful feature information since it only considers the pixel features of the picture centre and its neighbourhood ignoring the difference in amplitude.

2.2.3. ASM/AAM

The Active Shape Model (ASM) proposed in [49] is based on statistical models and is generally used to extract feature points on expression contours. This model mainly uses the global shape model to match the initial shape of the human face, and then establish a local texture model to obtain the contour features of the target more accurately. The Active Appearance Model (AAM) [50] is developed on the basis of ASM by incorporating local texture features.

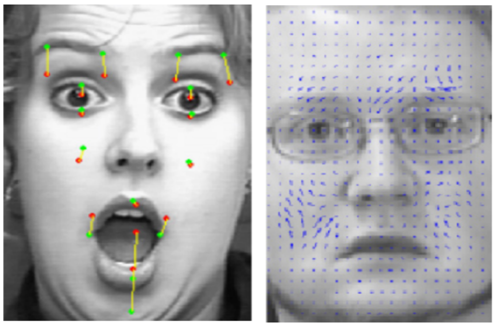
Cristinacce [51] fuses PRFR (Pairwise Reinforcement of Feature Responses) with AAM to detect feature points of local edges such as facial organs. Saatci [52] subtly cascades AAM with the SVM classifier to improve the recognition rates.

Despite ASM is more efficient than AAM, AAM can obtain higher recognition rate than ASM as it can better fit texture features.

2.2.4. Optical Flow Method

Optical flow is the pattern of apparent motion caused by the relative motion. The features of the continuous moving face image sequence are extracted by using Horn–Schunck (HS) optical flow [53] to combine the two-dimensional velocity field and the grayscale.

An approach for analysing and representing facial dynamics is presented in [54]. The algorithm computes optical flow caused by facial expressions to identify the direction of motions. In [55], an optical flow-based approach is developed and implemented to capture emotional expression by automatically recognising subtle changes in facial expressions. Sánchez et al. [56] systematically compare two optical flow-based FER methods, respectively referred to as feature point tracking and dense flow tracking, as shown in Figure 6.



(**a**) feature point tracking (**b**) dense flow tracking

Figure 6. Applications of optical flow-based methods on facial images in [56].

2.2.5. Haar-like Feature Extraction

The Haar-like feature template [57] is the combination of edge, linear, centre and diagonal features. Since these features all follow the step function proposed by Alfred Haar, they are called Haar-like. The feature template is divided into two rectangle regions, i.e., white and black, and the template's feature values are defined as the differences between value of white rectangle pixels and black rectangle pixels. The Haar eigenvalue reflects the grayscale variation of the image.

In [58], in order to represent the temporal variations in appearance of human face, dynamic Haar-like features are defined and encoded into binary pattern features.

When the global region illumination is stable, Haar can extract more facial motion unit change information because it describes the local grayscale variation of the face.

2.2.6. Feature Point Tracking

The main purpose of feature point tracking method is to synthesise the input emotional expressions according to the displacement of the feature points, as presented in Figure 7.

Tie et al. [59] extract over 20 points from the video stream as the feature points of the face model, then a variable 3D expression recognition model is constructed by tracking these feature points with particle filters. Liu et al. [60] propose a feature point tracking method based on Kanade-Lucas-Tomasi (KLT) and Scale Invariant Feature Transform (SIFT). The algorithm improves the SIFT, making the feature points evenly distributed without aggregation. Then, the KLT matching algorithm is hierarchically iteratively designed, so that the match can be quickly tracked when the target has obvious attitude and size change.

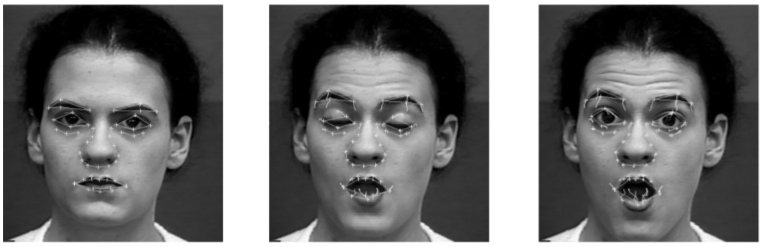


Figure 7. Feature points displacement.

Feature extraction is a decisive phase in FER. Note that some classifiers may pose the problem of curse of dimensionality (i.e., the phenomenon that data becomes sparser in high-dimensional space) and over-fitting when the number of extracted features is overwhelming. Dimensionality reduction methods are frequently embedded in this scenario, which can improves learning performance, increase computational efficiency, and decrease memory storage. Traditional feature selection methods including PCA and LDA have broad applications in many machine learning tasks. Feature value selection designed for outlier detection or object recognition is proposed in [61], which can be employed in the data with binary or nominal features.

2.3. Expression Classification

Another key to affecting the expression recognition rate is how to select the appropriate classifier that can successfully predict the face expressions. Commonly used and widely applied classifier in FER systems include kNN (k-Nearest Neighbours), SVM (Support Vector Machine), Adaboost (Adaptive Boosting), Bayesian, SRC (Sparse Representation-based Classifier), and PNN (Probabilistic Neural Network). Their merits and limitations, their comparisons are as below.

The kNN algorithm [62–64] is simple and easy to imply. The training speed of the algorithm is slow because every added new sample must be compared with the training set. Another notable characteristic of the kNN algorithm is that it is sensitive to the local structure of the data. The same weight of each attribute may result in suboptimal and unstable classification accuracy.

The SVM [41,65–69] can find a good compromising solution on complex models by providing limited sample data information to obtain generalisation ability. It is also possible to map linearly indivisible data to higher dimensions by kernel functions to convert the data into linear separable. By introducing a kernel function, the computer can effectively process high-dimensional data, and avoid dimension disasters to some extent.

AdaBoost [70–73] is sensitive to noisy and anomaly data. In some problems, it can be less susceptible to the overfitting problem than other learning algorithms. AdaBoost (with decision trees as the weak learners) is often referred to as the best out-of-the-box classifier.

Naive Bayes classifier [74–76] is highly scalable, requiring linear parameters for the number of variables in learning problems. One advantage is that only a small amount of training data is needed to estimate the parameters required for classification.

SRC [40,47,77,78] has better recognition effect than the traditional method, especially when the sample is subjected to random pixel corruption or random block occlusion. However, when handling data having the same direction distribution, SRC may not classify the data, since the sample vectors of different classes are distributed on the same vector direction.

As a classifier, PNN has fast training process and a structure of inherently parallel, to ensure convergence to the optimal classifier with the size of the representative training set increases [79]. Some implements in FER system are proposed in [80–82].

The conventional FER approaches is obviously less dependent on data and hardware compared to deep-learning-based approaches. However, feature extraction and classification have to be designed manually and separately, which means these two phases cannot be optimised simultaneously. The effectiveness of conventional FER methods is bound by the performance of each individual component.

3. Deep Learning-Based FER Approaches

Deep learning has demonstrated outstanding performance in many machine learning tasks including identification, classification, and target detection. In terms of FER, deep learning-based approaches highly reduce the reliance on image preprocessing and feature extraction and are more robust to the environments with different elements, e.g., illumination and occlusion, which means that they can greatly outperform the conventional approaches. In addition, it has potential capability to handle high volume data.

3.1. Convolutional Neural Network (CNN)

CNN [83] is an "end-to-end" model, an improvement of the Artificial Neural Network (ANN). The traits of CNN include local connectivity and weight sharing, resulting in less network parameters, faster training speed, and regularisation effect. An example of CNN-based FER procedure is shown in Figure 8.

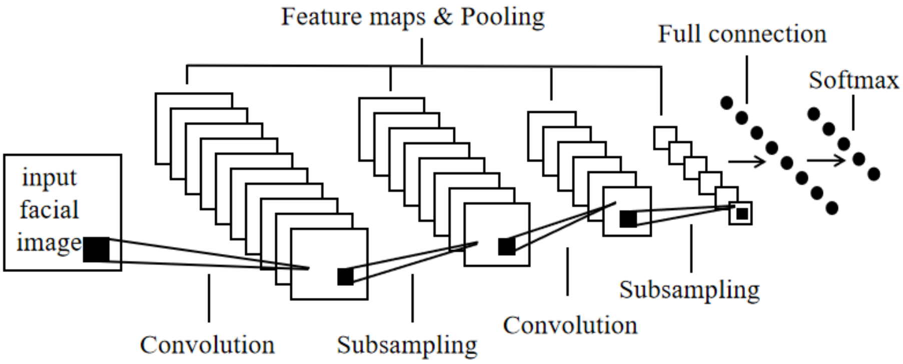


Figure 8. Procedure of CNN-based FER approach.

CNN is directly adapted for AU detection in most of the deep learning-based FER approaches. The CNN-based FER method proposed in [84] employs Facial Action Coding System (FACS) feature representation, which shows a generalisation capability of networks both cross-data and cross-task related to FER. A well-performed recognition rate is obtained, when utilising the model to the task of micro-expression detection.

Liu et al. [85] proposed a deformable facial action parts model for dynamic expression analysis. A deformable facial parts learning module is incorporated into the 3D CNN that can detect a particular facial action part under structured spatial constraint, and simultaneously obtain a representation based on the discriminating part.

In order to take advantage of temporal features for recognising facial expression, a new integration method named Deep Temporal Appearance-Geometry Network (DTAGN) is proposed in [86]. The Deep Temporal Appearance Network (DTAN) extracts appearance features from image sequences. The Deep Temporal Geometry Network (DTGN) extracts geometry features from temporal FL points. The the joint DTAGN boost the performance of the FER by making full use of temporal information.

The proposed network in [87] employs two convolutional layers and four inception layers to extend both the depth and width of the network and keep the computational budget constant as well. Compared with conventional CNN approaches, the proposed approach has shallower networks and has clear advantage in classification accuracy on cross-database evaluation scenarios.

The work [88] is about occlusion aware facial expression recognition. Considering different facial ROIs (Regions of Interest), Li et al. introduce two versions of ACNN (Convolution Neutral Network with Attention mechanism), i.e., pACNN (Patch based ACNN) and gACNN (global–local based ACNN). The pACNN focuses on local facial patches, while the gACNN combines local patches with global images. Experimental results show that ACNNs are able to improve the FER accuracy on occluded facial images by replacing the occluded patches with other related but non-occluded patches.

3.2. Deep Belief Network (DBN)

The DBN [89] is based on Restricted Boltzmann Machine (RBM) [90] and its feature extraction of input signal is unsupervised and abstract. The FER method based on DBN can learn the abstract information of facial images automatically and is susceptible to activity factors. Combined with other components, DBN has proved to be an effective FER approach.

A novel Boosted Deep Belief Network (BDBN) [91] is presented to combine feature learning/strengthen, feature selection, and classifier construction in a unified framework by performing three training stages iteratively. A set of features are fine-tuned jointly through this BDBN framework, and then selected to form a strong classifier. Additionally, based on their relative significance to the strong classifier, the discriminative properties of selected features are iteratively enhanced and then learn highly complex features from the facial image.

Zhao et al. [92] propose a new method of FER is proposed (denoted by DBNs + MLP) by integrating the DBN as the unsupervised feature learning module with the MLP (Multi-Layer Perceptron) as the classification module. First, the DBN is used to obtain abstract features from the primary pixels of facial images. Then, the MLP model is initialised to perform the classification processing with the aid of DBN's learning results.

An FER model combining LBP/VAR and DBN is given in [93]. The LBP/VAR feature, which is robust to light and rotation, is first extracted as the preprocessing of the model. Then the DBN is implemented for the second feature extraction process and expression classification.

In [94], the robust LDPP-PCA-GDA (Local Directional Position Pattern, Principal Component Analysis, Generalised Discriminant Analysis) features are combined with DBN for modelling the expressions as well as recognition. The proposed method consists of tolerance against illumination variation and extracts salient features by considering the highest strength directional position and the signs of the strengths. Furthermore, the recognition performance was superior over the traditional approaches.

3.3. Long Short-Term Memory (LSTM)

An RNN composed of LSTM units is commonly referred to as an LSTM network, which is well suited for the temporal features extraction of consecutive frames. Based on previous studies, long-range context modelling helps improve the accuracy of emotion analysis, some LSTM-based FER approaches on video sequence are proposed.

An LSTM-based system is built in [95] to assess dimensional representation of emotions in audiovisual scenarios. So as to get closer to the human performance when judging emotions, acoustic, linguistic, and visual information are taken into consideration that can reflect the realities of natural interactions.

Kim et al. [96] take advantage of CNN to learn spatial features of representative state frames, and then the LSTM of a facial expression is employed to learn the temporal features of the spatial feature representation. The proposed method utilised representative expression states in the network training which are determined in facial sequences in spite of the intensity of expression or the duration of expression.

The 3D Convolutional Neural Network (3DCNN) architecture in [97], combined the 3D Inception-ResNet (3DIR) layers with an LSTM unit. The extended Inception-ResNet module extracts the spatial relations within expression images. LSTM takes these temporal relations into consideration and applies this information to classify the sequences. Facial landmarks, visual highlights of facial components, are used as inputs to our network.

3.4. Generative Adversarial Network(GAN)

GAN [98] is an unsupervised learning model composed of a generative network and a discriminative network, which has been successfully adopted for image synthesis to generate facial images, videos and other impressively realistic images. GAN-based models not only conducive to training data augmentation and the corresponding recognition tasks, but also for pose-invariant and identity-invariant expression recognition. Figure 9 shows some pose-invariant facial images.

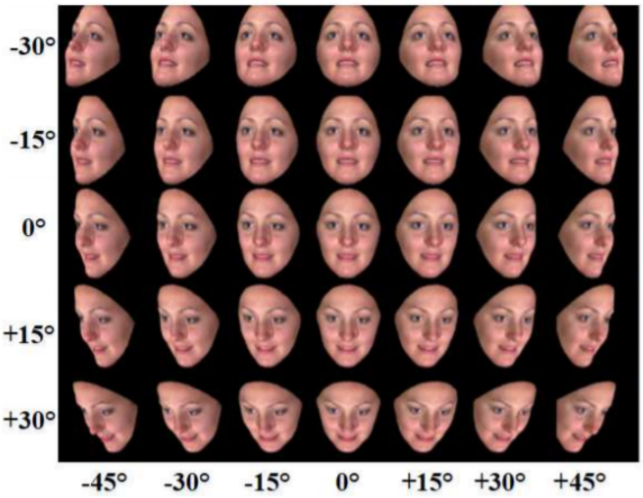


Figure 9. Facial images with different poses.

For multi-view FER, Lai et al. propose a multi-task GAN-based learning approach [99], where the generator frontalises a frontal face image based on the input non-frontal image while retaining the identity and expression information, and the discriminator is trained to distinguish and recognise. This face frontalisation system is demonstrated as valid for FER with visible head pose variations.

An end-to-end GAN-based model is presented in [100]. The encoder–decoder structure of the generator first learn a identity representation for face images which is then demonstratively specified through the expression and pose codes. In addition, the model is able to enlarge the FER training set by automatically generate face images with arbitrary expressions and head-poses.

In [101], GANs are employed to train the generator to generate six basic expressions from a face image while CNN is fine-tuned for each single identity sub-space expression classification. This model can alleviate the effect of inter-subject variations, and is adaptable to integrate with other CNN framework for FER.

To protect the user's privacy, Chen et al. [102] present a Representation-Learning Variational Generative Adversarial Network (PPRL-VGAN) to learn an image representation that is demonstratively specified from the identity information. A DeRL (De-expression Residue Learning) procedure is proposed in [103] to extract information of the expressive component. A cGAN (conditional Generative Adversarial Networks) generates the relevant neutral face image for any input to filter out the expressive information.

4. Datasets

Training and testing on existing datasets is a frequently used method of expression recognition. In this section, some FER-related datasets are provided and discussed with data augmentation methods.

4.1. Basic Information

Two-D static images are commonly used in the early studies; 2D video sequences are then used for FER studies because they promote recognition of expression with various dimensions. However, it is difficult to analyse facial deep information based on 2D data. The effectiveness of FER can be greatly influenced by some factors such as posture and light. Some 3D-based datasets are further used to handle micro facial behaviours and the variations of facial structure. In order to meet the needs of practical applications appropriately, some researchers aim to use datasets from "in the wild" rather than "in the lab". In addition, some datasets are designed for individual facial expression, which can be better used for the training of the corresponding classifiers.

Some popular datasets related to FER are introduced with basic information in this sub-section. Some sample images are shown in Figure 10.

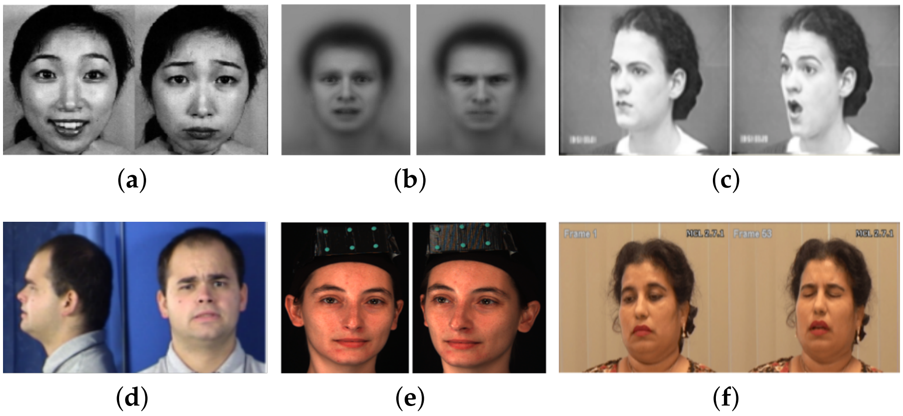


Figure 10. Examples of six representative datasets related to FER. (**a**) JAFFE [14]; (**b**) KDEF [104]; (**c**) CK+ [15]; (**d**) MMI [105]; (**e**) MPI [106]; (**f**) UNBC [107].

4.1.1. Japanese Female Facial Expressions (JAFFE)

JAFFE dataset [14] contains 213 images of seven facial emotion (six basic emotions and one neutral) posed by ten Japanese females. Each of the images is rated based on six emotion adjectives using 60 Japanese subjects. The original images are 256 pixels × 256 pixels.

4.1.2. Extended Cohn–Kanade Dataset (CK+)

CK+ dataset [15] is an extension of the CK dataset, containing 593 video sequences and still images of seven facial emotions (6 basic emotions and one contempt). The still images are posed in a lab situation, and the videos are shot in a similar situation. The age of 123 subjects range from 18 to 30. The resolution of the image are 640 pixels × 480 pixels and 640 pixels × 490 pixels, and the grey value is 8-bit precision.

4.1.3. Compound Emotion Dataset (CE)

CE dataset [21] contains 5060 images, containing 22 BEs and CEs of 230 subjects with an average age of 23, including many races. Facial occlusion is minimised, without glasses. Male subjects are required to be shaved, and all subjects are also asked to fully reveal the eyebrows. They are colour images with 3000 pixels × 4000 pixels resolution.

4.1.4. Denver Intensity of Spontaneous Facial Action Dataset (DISFA)

DISFA dataset [108] contains 130,000 stereo videos of 27 subjects with different gender and ethnicities. The images are acquired at high resolution (1024 pixels × 768 pixels), and all video frames are manually scored with the intensity of AU's (0–5 scale). A total of 66 facial landmarks of each images in the dataset are also labelled.

4.1.5. MMI Facial Expression Dataset

The MMI Fcial Expression dataset [105] include over 2900 videos and high-resolution still images of 75 subjects. Each of the AUs in videos are completely annotated. The original face images size are 720 pixels × 576 pixels.

4.1.6. Binghamton University 3D Facial Expression (BU-3DFE)

BU-3DFE [109] is designed for the research of 3D faces and facial expressions, as well as the development of a general understanding of human behaviour. It contains 100 subjects (56 females and 44 males), with a variety of ethnic/racial ancestries. The age range is between 18 and 70. The dataset include six emotions. There are 25 3D facial sentiment models for each subject in the dataset and a set of 83 manually annotated facial landmarks associated with each model. The original size of each face image is 1040 pixels × 1329 pixels.

4.1.7. Binghamton-Pittsburgh 3D Dynamic Spontaneous (BP4D-Spontaneous)

BP4D-spontaneous [110] is structured by 41 participants (23 females and 18 males, from 18 to 29 years old). An emotion elicitation protocol is designed to elicit emotions of participants effectively. Eight tasks are covered with an interview process and a series of activities to elicit eight emotions. For each task, there are both 3D and 2D videos. Meanwhile, the meta-data include manually annotated action units (AU), automatically tracked head pose, and 2D/3D facial landmarks. The original size of each face image is 1040 pixels × 1329 pixels.

4.1.8. Large MPI Facial Expression Database (MPI)

The MPI Facial Expression Database [106] is a validated database of emotional and conversational facial expressions. The dataset contains 55 different facial expressions performed by 19 German participants (ten females and nine males). Expressions are elicited with the help of a method-acting protocol, which guarantees both well-defined and natural facial expressions. All facial expressions are available in three repetitions, in two intensities, and from three different shooting angles. A detailed frame annotation is provided, from which a dynamic and a static version of the dataset are created. The results of an experiment are presented with two conditions that serve to validate the naturalness and recognisability of the video sequences, and the context scenarios.

4.1.9. Karolinska Directed Emotional Face (KDEF)

The KDEF dataset [104] involves 4900 images of human emotional facial expressions. The dataset consists of 70 people, each showing seven different emotional expressions taken from five different angles. The original size of each face image is 562 pixels × 762 pixels.

4.1.10. NVIE Dataset

NVIE [111] is a natural visible and infrared facial expression dataset, which contains both spontaneous and posed expressions of more than 100 subjects, with illumination provided from three different directions. The posed dataset includes the apex expressional images with and without glasses. It is labelled with six facial emotions, expression intensity, and Arousal–Valence label.

4.1.11. CMU Multi-PIE Database (Multi-PIE)

The CMU Multi-PIE database [112] is used for research in face recognition across pose and illumination. It contains 337 subjects, captured under 15 view points and 19 illumination conditions in four recording sessions for a total of more than 750,000 images. The labels are AAM-style with between 39 and 68 feature points.

4.1.12. Oulu-CASIA NIR-VIS Database (Oulu-CASIA)

The Oulu-CASIA NIR-VIS database [113] consists of 2880 image sequences with six basic expressions from 80 people between 23 and 58 years old. All expressions are captured in frontal direction with three different illumination conditions: normal, weak and dark. Subjects were asked to make a facial expression according to an expression example shown in picture sequences. The imaging hardware works at rate of 25 frames per second and image resolution is 320 × 240 pixels.

4.1.13. FER2013 Face Dataset

The FER2013 dataset [114] is actually provided for a competition of Kaggle. The dataset contains 35,887 face images, including 28,709 training sets, 3589 verification sets, and, 589 test sets, all of which are grayscale images of 48 pixels × 48 pixels. These samples are divided into seven categories on a basically average distribution, i.e., angry, disgusting, fearful, happy, neutral, sad, and amazed. Each sample in the dataset has a large difference in age, facial direction or other aspects, which is close to the real world situation.

4.1.14. GEMEP-FERA

The GEneva Multimodal Emotion Portrayals (GEMEP) [115] is a collection of audio and video recordings featuring ten actors portraying 18 emotional states, with different verbal contents and different modes of expression. This corpus consists of more than 7000 audio-video emotion portrayals, representing 18 emotions (including rarely studied subtle emotions), portrayed by ten professional actors who are coached by a professional director.

4.1.15. Acted Facial Expressions in the Wild Dataset (AFEW)

AFEW [116] is a dynamic temporal facial expressions data corpus with spontaneous expressions, various head poses, occlusions and illuminations, which is close to real world environment. Samples are labeled with seven categories of emotions (six basic emotion and one neutral). AFEW is divided into three data partitions in an independent manner in terms of subject and movie/TV source: Train (773 samples), Val (383 samples) and Test (653 samples), which ensures data in the three sets belong to mutually exclusive movies and actors.

Static Facial Expressions in the Wild Dataset (SFEW) [117] is developed by selecting frames from AFEW, has been divided into three sets: Train (958 samples), Val (436 samples) and Test (372 samples). It is labelled with seven facial expressions (six basic emotion and 1 neutral).

4.1.16. Real-world Affective Database (RAF-DB)

RAF-DB [118,119] is a large-scale facial expression database with 29,672 number of real-world facial images. Each image has been independently labelled by about 40 annotators with seven classes of basic emotions and 12 classes of compound emotions, five accurate landmark locations, 37 automatic landmark locations. Images in this database are of great variability in the subjects' age range, gender, ethnicity attributes, head poses, lighting conditions, occlusions, and post-processing operations, etc.

Real-world Affective Faces Multi Label (RAF-ML) [120] is a multi-label facial expression dataset. 4908 number of real-world images with blended emotions are provided with 6-dimensional expression distribution vector and landmark locations per image. Specially, 315 well-trained annotators are employed to ensure each image can be annotated enough independent times.

4.1.17. GENKI-4K Dataset

The MPLab GENKI dataset [121] is an extended dataset of images containing faces spanning a wide range of illumination conditions, geographical locations, personal identity, and ethnicity. The images in the dataset are divided into overlapping subsets, each with its own labels and descriptions. The GENKI-4K subset is dedicated for smile recognition, containing 4000 face images labelled as either "smiling" or "non-smiling" and their head-pose by human coders.

4.1.18. The UNBC-Mc Master Shoulder Pain Expression Archive Dataset

This dataset [107] is available to accelerate research into pain and augment current datasets. It includes 200 video sequences including spontaneous facial expressions, 48,398 FACS coded frames, associated pain frame-by-frame scores and sequence-level self-report and observer measures, and 66-point AAM landmarks.

4.2. Discussion

According to the peculiarity of human facial expression, a dataset has four notable elements, namely image dimension, shooting environment, labelling method, and elicitation method. An overview of FER-related datasets is presented in Table 2.

4.2.1. Image Dimension

Human facial expression images can be divided into 2D and 3D according to the dimensionality.

- **2D-type:** The traditional 2D laboratory dataset usually has good separability of different categories, due to its exaggerated expression and limited variables. The JAFFE dataset [14] specially uses Japanese females as the subjects. CE [21] consists of 22 categories of emotions of 230 subjects with facial occlusion minimised. This type of dataset is useful for understanding the procedure of expression recognition and comparing the performances of different experimental methods.
- **3D-type:** Establishment of 3D-based facial expression dataset accelerates the examination of 3D spatiotemporal features in subtle facial expression, revealing the connection between pose and motion dynamics in AUs, and better acquainting of spontaneous facial action. The BU-3DFE [109] and BP4D-Spontaneous [110] are published to accelerate the research on the facial behaviour and 3D structure of facial expressions.

4.2.2. Shooting Environment

The shooting environment is related to the quality and capacity of image data.

- **Unique condition:** Positive face images under a single specific condition can provide accurate feature information for expression recognition, but the trained model is very limited in scope, e.g., JAFFE [14] and CK+ [15].
- **Complex condition:** In order to improve the application scope and processing ability of the FER model, some datasets selectively collect expression images of various illumination conditions, face directions and head postures, e.g., MPI [106] and Oulu-CASIA [113].
- **Wild condition:** Facial expressions in the wild is close to real world environment, e.g., AFEW [116] extracted from movies and RAF-DB [118,119] downloaded form the Internet. Such datasets are more challenging.

Table 2. An overview of FER-related datasets.

Dataset	Subjects	Samples	Resolution	Condition	Elicitation	Annotation	Source
JAFFE [14]	10	213 still images	256 × 256	unique	posed	6 BEs & neutral	http://www.kasrl.org/jaffe.html
CK+ [15]	123	593 still images & video sequence	640 × 480, 640 × 490	unique	posed	6 BEs & contempt	http://www.pitt.edu/~emotion/ck-spread.htm
CE [21]	230	5060 still images	3000 × 4000	unique	posed	22 CEs	http://cbcsl.ece.ohio-state.edu/dbform_compound.html
DISFA [108]	27	130,000 stereo videos	1024 × 768	unique	posed	AU's intensity scale, 66 FLs	http://www.engr.du.edu/mmahoor/DISFA.htm
MMI [105]	75	740 still images & 2900 video sequence	720 × 576	complex	posed	6 BEs & neutral, AUs	http://mmifacedb.eu/
BU-3DFE [109]	100	2500 still images	1040 × 1329	complex	posed	6 BEs, 83 FLs	http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html
BP4D-Spontaneous [110]	41	328 videos	1040 × 1329	unique	spontaneous	8 expressions, AUs, FLs	http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html
MPI [106]	19	1045 still images & video sequence	768 × 576	complex	spontaneous	55 expressions	https://www.b-tu.de/en/graphic-systems/databases/the-small-mpi-facial-expression-database
KDEF [104]	70	4900 still images	562 × 762	complex	posed	6 BEs & neutral	http://www.emotionlab.se/resources/kdef
Multi-PIE [112]	337	755,370 still images	N/A	complex	posed	6 expressions, FLs	http://www.flintbox.com/public/project/4742/
Oulu-CASIA [113]	80	2880 image sequence	320 × 240	complex	posed	6 BEs	http://www.cse.oulu.fi/CMV/Downloads/Oulu-CASIA
FER2013 [114]	N/A	35,887 still images	48 × 48	wild	posed & spontaneous	6 BEs & neutral	https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data
GEMEP-FERA [115]	10	7000 audio-video	N/A	wild	spontaneous	18 expressions	https://gemep-db.sspnet.eu/
AFEW [116]	N/A	1809 videos	N/A	wild	posed & spontaneous	6 BEs & neutral	https://cs.anu.edu.au/few/AFEW.html
SFEW [117]	N/A	1766 still images	N/A	wild	posed & spontaneous	6 BEs & neutral	https://cs.anu.edu.au/few/AFEW.html
RAF-DB [118,119]	N/A	29,672 still images	N/A	wild	posed & spontaneous	6 BEs & neutral, 12 CEs, 42 FLs	http://www.whdeng.cn/RAF/model1.html
RAF-ML [120]	N/A	4908 still images	N/A	wild	posed & spontaneous	6 BEs distribution vector, 42 FLs	http://www.whdeng.cn/RAF/model2.html
GENKI-4K [121]	N/A	4000 still images	N/A	wild	spontaneous	(smiling & non-smiling), head-pose	http://mplab.ucsd.edu/wordpress/?page_id=398
UNBC [107]	N/A	200 video sequence & 48,398 frames	N/A	wild	spontaneous	66 FLs, AUs, expression intensity	http://www.pitt.edu/~emotion/um-spread.htm

4.2.3. Annotation Method

Facial expression datasets are usually labelled with the expression categories directly. Such as JAFFE [14], CK+ [15], and FER2013 [114] are annotated with 7 expression categories. Datasets of one particular category of expression, e.g., GENKI-4K [121] labelled as "smiling" and "non-smiling", benefit to better understand the specific kind expression on different dimensions and degrees. In addition, some datasets give one or more annotation information in AUs, FLs, and intensity for more diverse research. Taking the DISFA dataset [108] as an example, 66 FLs and the intensity of AU's are labelled in each images.

4.2.4. Elicitation Method

For FER datasets, the induction method can be simply defined as "posed" and "non-posed" (spontaneous). Posed expression datasets are often exaggerated, magnifying the differences between categories and making them easy to be classified. Spontaneous expressions are elicited under the guarantee that both are well-defined and natural, better reflecting the real world.

4.3. Data Augmentation

In real-world applications of image recognition, training data is normally very limited, which is generally the main reason for the over-fitting problem and sub-optimal accuracy. On the contrary, using more training data is the common method to obtain better performance. Note that utilising overwhelming training data can lead to low efficiency and excessive consumption of computing resource. However, it is unrealistic to obtain sufficient training samples in real-world applications. Therefore, we can augment the training data artificially, which is a widespread preprocessing method in image recognition.

For image recognition, we can augment the training data by rotating, transforming, shearing, random scaling the image, etc.. Note that choosing an appropriate method for different applications is important, i.e., the augment methods need to conform to the possible changes in real world. Random cropping and mirroring are commonly used in FER, because they can better induce the spatial invariance of human faces and have an efficient performance in data expending. An example of data augmentation of human facial images is shown in Figure 11.

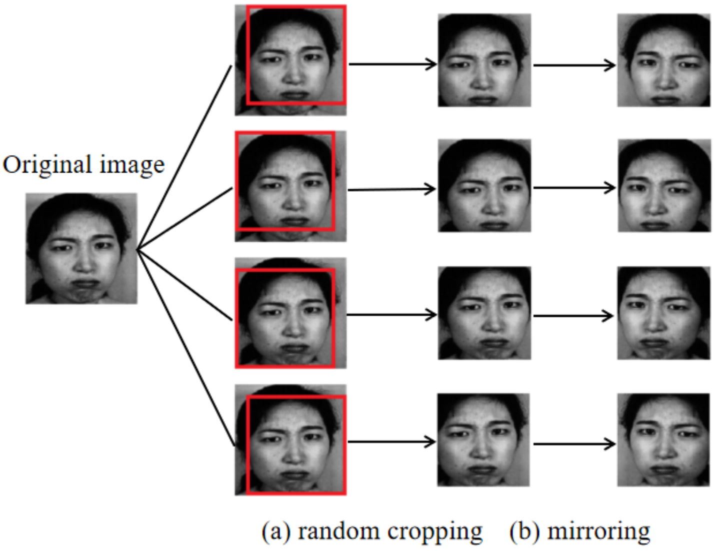


Figure 11. An example of data augmentation(face images are taken from JAFFE dataset [14]).

Jinwoo et al. [122] apply data augmentation both on training phase and testing phase. They use 42 pixels × 42 pixels randomly cropped images and their mirrored images to attain eight times more data for training input. An averaging method is then used on the testing phase to reduce outliers. The probability is averaged as the final output of their cropped images and mirrored images.

5. Performance Metrics

In practical tasks, multiple learning algorithms can be chosen, and even for the same learning algorithm, different parameters lead to a variety of results. Evaluation metrics are critical to identify the merits of a method because it provides a standard for quantitative comparisons. In this section, we present the evaluation methods and evaluation metrics that are publicly available in the FER studies. The recognition rate of different methods is also compared with the FER typical classification method introduced in the previous section.

5.1. Evaluation Methods

The difference among various evaluation methods is in dividing the samples into training sets and test sets. In such multi-classification tasks like FER, each category of emotion should be divided into training sets and test sets in the same way, and the model is evaluated according to the average performance of each category of emotion. Commonly used evaluation methods include the hold-out method, K-fold cross-validation, leave-one-out cross-validation (LOOCV), and bootstrapping method.

The hold-out method can avoid over-fitting since the test set and the training set are separated. One shortcoming of the hold-out method is that the evaluation result is sensitive to the ratio of the training set and the validation set partition. The K-fold cross-validation makes full use of all the samples, and avoids over-fitting and under-fitting effectively, but the computational complexity of this method is related to the parameter k, i.e., it needs to be trained k times and tested k times. LOOCV is a special case of K-fold cross-validation when the value of parameter k is equal to the number of the samples. LOOCV is suitable for small samples because the sample utilisation rate is the highest. Nevertheless, high utilisation will lead to high computational complexity when dealing with large sample problems.The bootstrapping method is useful when the sample size is small and it is difficult to partition the training set and the test set effectively. However, the bootstrapping method changes the distribution of the initial data, which introduces an estimated bias.

5.2. Evaluation Metrics

The evaluation metric plays a vital role during the training process and the selection of which is an important key for discriminating and obtaining the optimal classifier. FER is naturally a multi-class classification problem, Acc (accuracy), i.e., the proportion of the samples that are correctly classified is a direct performance evaluation metric. In order to comprehensively take the recognition effect for each category of expression into consideration, the final accuracy can also be defined as the average of the recognition accuracy of each category of expression. These two methods of accuracy calculation are called overall accuracy and average accuracy, respectively. In general, higher accuracy stands for better classification performance.

The definition of Acc is given below:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

Some metrics for binary classification can be extended for multi-class classification evaluations [123] (e.g., Precision P, Recall R, and F-measure). Apart from the metrics related to effectiveness, other evaluation methods for measuring the efficiency and scalability of the classifier, e.g., execution time, training time, and resource occupancy, also need to be considered in practice.

5.3. Evaluation Results and Discussion

In this section, we perform some of the aforementioned methods on the benchmark datasets to illustrate the performance of existing state-of-art FER approaches, as shown in Table 3. In the next part, discussions of the evaluation results are presented from two aspects respectively.

Table 3. Comparison of Representative FER Approaches on Widely Evaluated Datasets.

Database	Approaches	Accuracy	Database	Approaches	Accuracy
JAFFE [14]	[40] Gabor + SRC	88.57	MMI [105]	[47] LPQ + SRC	62.72
	[41] Gabor + SVM	80.95		[77] LBP + SRC	59.18
	[44] LBP + LP	93.80		[85] CNN (3DCNN-DAP)	63.40
	[46] LBP (LDP)	90.10		[86] CNN (DTAJN)	70.24
	[62] KNN	90.76		[87] CNN (DeeperCNN)	77.90
	[63] PCA + FSVM/KNN	87.70		[88] CNN (ACNN)	70.37
	[66] SVM	97.10		[96] CNN + LSTM	78.61
	[69] IDA + SVM	92.73		[97] 3DIR + LSTM	79.26
	[71] Haar + Adaboost	98.90		[103] GAN (PPRL-VGAN)	73.23
	[78] LBP/Gabor + SRC	84.76	FERA [115]	[87] CNN (DeeperCNN)	76.70
	[81] AAM + PNN	96.00		[97] 3DIR + LSTM	77.42
	[91] DBN (BDBN)	91.80	FER2013 [114]	[68] Cubic SVM+HoG	57.17
	[92] DBN + MLP	90.95		[84] CNN	72.10
CK+ [15]	[42] Gabor (Gabor-mean-DWT)	92.50		[87] CNN(DeeperCNN)	61.10
	[46] LBP (LDP)	96.40	BU-3DFE [109]	[75] Bayesian	80.47
	[56] Optical Flow	95.45		[99] GAN	73.13
	[67] ASM + SVM	94.70		[100] GAN	81.20
	[71] HoG + Adaboost	88.90		[101] GAN (IA-gen) + CNN	78.83
	[78] LBP/Gabor + SRC	97.14		[103] GAN (PPRL-VGAN)	84.17
	[82] Gabor + PNN	89.00	Multi-PIE [112]	[75] Bayesian	90.24
	[84] CNN	98.62		[87] CNN (Deeper CNN)	94.70
	[85] CNN (3DCNN-DAP)	92.40		[99] GAN	87.08
	[86] CNN (DTAJN)	97.25		[100] GAN	91.80
	[87] CNN (Deeper CNN)	93.20	SFEW [117]	[75] Bayesian	44.72
	[88] CNN(ACNN)	91.64		[87] CNN(DeeperCNN)	47.70
	[91] DBN (BDBN)	96.70		[88] CNN (ACNN)	51.72
	[92] DBN + MLP	98.57		[100] GAN	26.58
	[93] LBP/VAR + DBN	91.40	Oulu-CASIA [113]	[88] CNN (ACNN)	58.18
	[97] 3DIR + LSTM	95.53		[101] GAN (IA-gen) + CNN	88.92
	[103] GAN (PPRL-VGAN)	97.30		[103] GAN (PPRL-VGAN)	88.00

5.3.1. Data Perspective Discussion

As illustrated in Table 3, the heterogeneity between datasets leads to disparate performance of FER approaches. A data perspective discussion of evaluation results is as below:

- A high accuracy of over 90% can be obtained by state-of-the-art FER approaches on JAFFE [14] and CK+ [15] datasets. As the early collections of pose-invariant human facial expressions with minimised facial occlusion, the majority of FER approaches are evaluated on JAFFE [14] and CK+ [15] datasets. And these two datasets are widely used as the benchmark for FER comparison.
- The accuracy of FER approaches on MMI [105] dataset is generally less than 80%. There are significant inter-personal variations, because of the subjects' non-uniformly performances and accessories (e.g., glasses, moustache). Experiments of eight FER methods illustrate that deep learning-based methods perform better on the MMI dataset, with about 10% higher accuracy.
- Unlike the three datasets above, the accuracy on Multi-PIE [112] and Oulu-CASIA [113] are about 80%. These two datasets are collected in complex conditions, which support the FER research for illumination variation. FER approaches are better performed on Multi-PIE even if the shooting conditions are more complicated (15 view points with 19 illumination condition in Multi-PIE, frontal direction with three illumination conditions in Oulu-CASIA). That is because the Multi-PIE is labelled with feature points as the a priori knowledge.
- The latest FER approaches, especially the GAN-based aproaches, can achieve about 80% accuracy on the BU-3DFE [109] dataset, which has different properties than other datasets. As a 3D-based facial expression dataset, BU-3DFE reveals the connection between pose and motion dynamics in

facial AUs, providing sufficient feature information for the study of multi-pose FER. Meanwhile, detailed 3D information and facial landmark annotation are applicable to GAN-based algorithm for image generation.

- The accuracy rates of FER on FERA [115], FER2013 [114], and SFEW [117] are approximately 70%, 60%, and 40%, respectively. These three datasets have more challenging conditions, i.e., subjects perform spontaneous expressions under the wild circumstance. Neither the conventional approaches nor the deep learning-based approaches have high accuracy in dealing with the FER in the wild condition, which is one of the challenges in this field.

5.3.2. Methodology Perspective Discussion

Through investigation, there are differences in the performance between conventional FER approaches and deep learning-based FER approaches, especially after wild datasets are proposed. The performances of these two approaches are discussed separately as below:

The conventional FER approaches are based on manual feature extraction and less dependent on data and hardware, which has advantages in small data sample analysis.

- Convention FER approaches can achieve promising results on 2D datasets collected under unique conditions. All the mentioned conventional FER approaches obtain a high accuracy of over 90% on the JAFFE [14] and CK+ [15] datasets, which are similar to deep learning-based FER approaches.
- Differences appear when FER approaches are applied to datasets that have significant interpersonal variations. Take the performance on the MMI dataset as an example, the average accuracy of the conventional FER approaches is about 60.94%, while the deep learning-based approaches is about 73.28%.
- For more challenging task like FER under wild environmental conditions, conventional approaches are rarely applied since feature extraction for complex datasets is still an obstacle.

Deep learning-based FER approaches highly reduce the reliance on image preprocessing and feature extraction and are more robust to environments with different elements. Here, we analyse the differences of four kinds of neural network according to the performance results:

- As shown in Table 3, CNN framework can be applied to almost all the FER datasets and achieve stable accuracy. The characteristics of CNN (e.g., local connectivity and weight sharing) make it advantageous in image processing.
- DBN is able to capture general appearance changes and to train on large but sparsely labelled datasets. By combining with other methods, DBN can achieve promising results for the FER on JAFFE [14] and CK+ [15] datasets. However, performance on other listed datasets currently cannot reveal its advantages well. Perhaps more attempts can be made for FER in wild environmental conditions in the future.
- Performances on MMI [105] and CK+ [15] datasets indicate that LSTM-based FER approaches have an advantage on video sequences. As a type of RNN that based on long-range context modelling, LSTM network is well suited for the temporal feature extraction of consecutive frames.
- Significant progress on BU-3DFE [109], Multi-PIE [112] and Oulu-CASIA [113] is made by using GAN-based FER aproaches. GAN model, composed of a generator and a discriminator, has been successfully applied in image synthesis to generate facial images, videos and other impressively realistic images. Hence, GAN-based models conducive for pose-invariant and identity-invariant expression recognition.

6. Challenges and Opportunities

Over the past decades, many attempts have been done in developing FER algorithms for both theoretical analysis and practical applications. As the FER literature shifts its main focus to the challenging wild environmental conditions, there are still several challenges and opportunities, which are mentioned and discussed in this section.

6.1. Wild Environmental Conditions

Complicated conditions like occlusion and pose-variation, which may hinder the recognition of original facial expressions, are two major obstacles for the adaptability of FER, especially in the wild scenarios.

Li et al. notice this issue, and introduce two versions of ACNN to replace the occluded patches with other related but non-occluded ones in [88]. Mao et al. [75] propose a hierarchical bayesian model for multi-pose FER. Additionally, the GAN has been employed in [99] to generate multi-view facial images with arbitrary expressions and poses for FER.

Even though some existing algorithms can process FER in special conditions, these methods can only adapt to limited variations and usually have low accuracy. The problem about how to improve the adaptability is significant to practically apply FER systems under the changeable and complex real-world environment.

6.2. The Lack of High-Quality Publicly Available Data

FER is a data-driven task. It generally requires a large amount of training data to capture subtle expression-related deformations when training a deep neural network. The major challenge that deep FER systems suffer is the shortage of training data in terms of both quantity and quality.

Some widely used FER-related datasets are introduced and discussed in Section 4, but the small scale of datasets is one of many common problems. Meanwhile, the problem of data bias and inconsistent annotations often exist when using different datasets due to diverse class distribution and the subjective annotation method. Another common problem is class imbalance. In-depth research and discussion on dataset bias and imbalanced distribution in expression recognition are conducted in [124,125].

An opportunity is to publish relatively extensive datasets that involve natural scenarios and include various facial expressions have recently become publicly available. Some significant achievements have been attained in datasets EmotioNet [19], AffectNet [126] and RAF-DB/RAF-ML [118–120].

6.3. The Pressure of High-Volume Data Processing

Available FER systems usually well perform on commonly used traditional datasets with small capacity and low pixel resolution. However, in many scientific and commercial applications, data is measured in terabytes, which poses challenges to the FER system in data storage, transmission and processing. Moreover, the requirement of data compression are urgent demanded when executing FER on real-time sequences. According to the FACS theory, expression related units are only part of the entire facial region. An idea worth trying for data compression is to reduce irrelevant facial information.

6.4. Multi-Modal Affect Recognition

Although individual FER according to visible facial images can obtain promising performance, combining with other models into an integrated system can provide additional information and further improve the reliability. Take, for instance, audio-based effect recognition proposed in [127], which can provide paralinguistic affective information in speech. Moreover, 3D facial images with depth information and infrared images with skin temporal record are not sensitive to illumination variations, which may be an efficient alternative for research of realistic facial expression. Additionally, facial expression can be further linked to the emotion states by the valence–arousal (VA) model [13], and then multi-modal dimensional emotion recognition [128–131] can be performed.

6.5. Visual Privacy

Growing privacy-preserving concerns are a major barrier in camera-equipped systems, such as the real-time FER for smartphones. Even if a few attempts have been made in [102,132,133], a wide variety proposed FER method generally depend on high-resolution images, but with little or no attention to protecting the visual privacy of their users. Hence, more trustworthy and accurate privacy protection methods are needed so as to strike a balance between privacy and data utility for FER systems.

7. Conclusions

Facial Expression Recognition (FER) has attracted increasing attention in recent years. The past decade has witnessed the development of many new FER algorithms. This paper provides a comprehensive review about recent advances in FER technology. We first introduce some related terminology and review the research background of FER. Then, we classify the existing FER methods into conventional methods and deep learning-based methods. In particular, we divide the conventional methods into three major steps, i.e., image preprocessing, feature extraction, and expression classification. In each step, various possible methods are introduced and discussed. In terms of deep learning-based methods, four kinds of popular deep learning networks are presented, and some related FER algorithms are reviewed and analysed. Besides, seventeen FER datasets are introduced. Four FER-related elements of datasets are subsequently summarised. In addition, some methods and metrics are given on how to evaluate these FER algorithms. At the end of the survey, we present some challenges and opportunities of the FER that require future research. This survey aims to provide an organised and detailed study of the work done in the area of FER and further promote the research in this field.

Author Contributions: Conceptualisation, Y.H.; writing–original draft preparation, Y.H.; writing–review and editing, Y.H., F.C. and X.W.; supervision, S.L. and X.W.; project administration, S.L.

Funding: This research was funded by the National Natural Science Foundation of China (61572513).

Acknowledgments: We would like to gratitude anonymous reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mehrabian, A.; Russell, J.A. *An Approach to Environmental Psychology*; The MIT Press: Cambridge, MA, USA, 1974.
2. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [CrossRef]
3. Bartlett, M.S.; Littlewort, G.; Fasel, I.; Movellan, J.R. Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. In Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop, Madison, WI, USA, 16–22 June 2003; Volume 5, pp. 53–53.
4. Bekele, E.; Zheng, Z.; Swanson, A.; Crittendon, J.; Warren, Z.; Sarkar, N. Understanding how adolescents with autism respond to facial expressions in virtual reality environments. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 711–720. [CrossRef] [PubMed]
5. Chen, C.H.; Lee, I.J.; Lin, L.Y. Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders. *Res. Dev. Disabil.* **2015**, *36*, 396–403. [CrossRef] [PubMed]
6. Assari, M.A.; Rahmati, M. Driver drowsiness detection using face expression recognition. In Proceedings of the IEEE International Conference on Signal and Image Processing Applications, Kuala Lumpur, Malaysia, 16–18 November 2011; pp. 337–341.
7. Jabon, M.; Bailenson, J.; Pontikakis, E.; Takayama, L.; Nass, C. Facial expression analysis for predicting unsafe driving behavior. *IEEE Perv. Comput.* **2011**, *10*, 84–95. [CrossRef]
8. Kapoor, A.; Burleson, W.; Picard, R.W. Automatic prediction of frustration. *Int. J. Hum.-Comput. Stud.* **2007**, *65*, 724–736. [CrossRef]
9. Lankes, M.; Riegler, S.; Weiss, A.; Mirlacher, T.; Pirker, M.; Tscheligi, M. Facial expressions as game input with different emotional feedback conditions. In Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology, Yokohama, Japan, 3–5 December 2008; pp. 253–256.

10. Jerritta, S.; Murugappan, M.; Nagarajan, R.; Wan, K. Physiological signals based human emotion recognition: A review. In Proceedings of the IEEE 7th International Colloquium on Signal Processing and its Applications, Penang, Malaysia, 4–6 March 2011; pp. 410–415.
11. Tian, Y.I.; Kanade, T.; Cohn, J.F. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 97–115. [CrossRef]
12. Russell, J.A. A circumplex model of affect. *J. Pers. Soc. Psychol.* **1980**, *39*, 1161. [CrossRef]
13. Chang, W.Y.; Hsu, S.H.; Chien, J.H. FATAUVA-Net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 17–25.
14. Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with gabor wavelets. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 200–205.
15. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
16. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. Facial landmark detection by deep multi-task learning. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin, Germany, 2014; pp. 94–108.
17. Wu, Y.; Ji, Q. Facial landmark detection: A literature survey. *Int. J. Comput. Vis.* **2019**, *127*, 115–142. [CrossRef]
18. Ekman, P.; Friesen, W.V. *Facial Action Coding System: Investigator's Guide*; Consulting Psychologists Press: Palo Alto, CA, USA, 1978.
19. Benitez-Quiroz, C.F.; Srinivasan, R.; Martinez, A.M. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5562–5570.
20. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [CrossRef]
21. Du, S.; Tao, Y.; Martinez, A.M. Compound facial expressions of emotion. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E1454–E1462. [CrossRef]
22. Ekman, P. Darwin, deception, and facial expression. *Ann. N. Y. Acad. Sci.* **2003**, *1000*, 205–221. [CrossRef] [PubMed]
23. Samal, A.; Iyengar, P.A. Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognit.* **1992**, *25*, 65–77. [CrossRef]
24. Fasel, B.; Luettin, J. Automatic facial expression analysis: A survey. *Pattern Recognit.* **2003**, *36*, 259–275. [CrossRef]
25. Sandbach, G.; Zafeiriou, S.; Pantic, M.; Yin, L. Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image Vis. Comput.* **2012**, *30*, 683–697. [CrossRef]
26. Danelakis, A.; Theoharis, T.; Pratikakis, I. A survey on facial expression recognition in 3D video sequences. *Multimed. Tools Appl.* **2015**, *74*, 5577–5615. [CrossRef]
27. Takalkar, M.; Xu, M.; Wu, Q.; Chaczko, Z. A survey: Facial micro-expression recognition. *Multimed. Tools Appl.* **2018**, *77*, 19301–19325. [CrossRef]
28. Kumari, J.; Rajesh, R.; Pooja, K. Facial expression recognition: A survey. *Procedia Comput. Sci.* **2015**, *58*, 486–491. [CrossRef]
29. Huang, D.; Shan, C.; Ardabilian, M.; Wang, Y.; Chen, L. Local binary patterns and its application to facial image analysis: A survey. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2011**, *41*, 765–781. [CrossRef]
30. Zhang, L.; Verma, B.; Tjondronegoro, D.; Chandran, V. Facial Expression Analysis under Partial Occlusion: A Survey. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 25. [CrossRef]
31. Deshmukh, S.; Patwardhan, M.; Mahajan, A. Survey on real-time facial expression recognition techniques. *IET Biometr.* **2016**, *5*, 155–163. [CrossRef]
32. Goyal, S.J.; Upadhyay, A.K.; Jadon, R.; Goyal, R. Real-Life Facial Expression Recognition Systems: A Review. In *Smart Computing and Informatics*; Springer: Berlin, Germany, 2018; pp. 311–331.
33. Khan, S.A.; Hussain, A.; Usman, M. Facial expression recognition on real world face images using intelligent techniques: A survey. *Opt.-Int. J. Light Electron Opt.* **2016**, *127*, 6195–6203. [CrossRef]

34. Viola, P.; Jones, M.J. Robust real-time face detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
35. Hsu, R.L.; Abdel-Mottaleb, M.; Jain, A.K. Face detection in color images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 696–706.
36. Shan, S.; Gao, W.; Cao, B.; Zhao, D. Illumination normalization for robust face recognition against varying lighting conditions. In Proceedings of the 2003 IEEE International SOI Conference, Nice, France, 17 October 2003; pp. 157–164.
37. Chen, W.; Er, M.J.; Wu, S. Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2006**, *36*, 458–466. [CrossRef]
38. Du, S.; Ward, R. Wavelet-based illumination normalization for face recognition. In Proceedings of the IEEE International Conference on Image Processing 2005, Genova, Italy, 14 September 2005; Volume 2, pp. II–954.
39. Tan, T.; Sim, K.; Tso, C.P. Image enhancement using background brightness preserving histogram equalisation. *Electron. Lett.* **2012**, *48*, 155–157. [CrossRef]
40. Zhang, S.; Li, L.; Zhao, Z. Facial expression recognition based on Gabor wavelets and sparse representation. In Proceedings of the IEEE 11th International Conference on Signal Processing, Beijing, China, 21–25 October 2012; Volume 2, pp. 816–819.
41. Yu, J.; Bhanu, B. Evolutionary feature synthesis for facial expression recognition. *Pattern Recognit. Lett.* **2006**, *27*, 1289–1298. [CrossRef]
42. Mattela, G.; Gupta, S.K. Facial Expression Recognition Using Gabor-Mean-DWT Feature Extraction Technique. In Proceedings of the 5th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 22–23 February 2018; pp. 575–580.
43. Ahonen, T.; Hadid, A.; Pietikäinen, M. Face recognition with local binary patterns. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2004; pp. 469–481.
44. Feng, X.; Pietikäinen, M.; Hadid, A. Facial expression recognition based on local binary patterns. *Pattern Recognit. Image Anal.* **2007**, *17*, 592–598. [CrossRef]
45. Guo, Z.; Zhang, L.; Zhang, D. A completed modeling of local binary pattern operator for texture classification. *IEEE Trans. Image Process.* **2010**, *19*, 1657–1663.
46. Jabid, T.; Kabir, M.H.; Chae, O. Robust facial expression recognition based on local directional pattern. *ETRI J.* **2010**, *32*, 784–794. [CrossRef]
47. Wang, Z.; Ying, Z. Facial expression recognition based on local phase quantization and sparse representation. In Proceedings of the 2012 8th International Conference on Natural Computation, Chongqing, China, 29–31 May 2012; pp. 222–225.
48. Chao, W.L.; Ding, J.J.; Liu, J.Z. Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection. *Signal Process.* **2015**, *117*, 1–10. [CrossRef]
49. Cootes, T.F.; Taylor, C.J.; Cooper, D.H.; Graham, J. Active shape models-their training and application. *Comput. Vis. Image Underst.* **1995**, *61*, 38–59. [CrossRef]
50. Cootes, T.F.; Edwards, G.J.; Taylor, C.J. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *6*, 681–685. [CrossRef]
51. Cristinacce, D.; Cootes, T.F.; Scott, I.M. A multi-stage approach to facial feature detection. In Proceedings of the British Machine Vision Conference (BMVC), Kingston, UK, 7–9 September 2004; Volume 1, pp. 277–286.
52. Saatci, Y.; Town, C. Cascaded classification of gender and facial expression using active appearance models. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 10–12 April 2006; pp. 393–398.
53. Horn, B.K.; Schunck, B.G. Determining optical flow. *Artif. Intell.* **1981**, *17*, 185–203. [CrossRef]
54. Yacoob, Y.; Davis, L.S. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Trans. Pattern Anal. Mach. Intell.* **1996**, *18*, 636–642. [CrossRef]
55. Cohn, J.F.; Zlochower, A.J.; Lien, J.J.; Kanade, T. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; p. 396.
56. Sánchez, A.; Ruiz, J.V.; Moreno, A.B.; Montemayor, A.S.; Hernández, J.; Pantrigo, J.J. Differential optical flow applied to automatic facial expression recognition. *Neurocomputing* **2011**, *74*, 1272–1282. [CrossRef]
57. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the CVPR, Kauai, HI, USA, 8–14 December 2001; Volume 1, pp. 511–518.

58. Yang, P.; Liu, Q.; Metaxas, D.N. Boosting encoded dynamic features for facial expression recognition. *Pattern Recognit. Lett.* **2009**, *30*, 132–139. [CrossRef]
59. Tie, Y.; Guan, L. A deformable 3-D facial expression model for dynamic human emotional state recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *23*, 142–157. [CrossRef]
60. Liu, Y.; Wang, J.D.; Li, P. A Feature Point Tracking Method Based on The Combination of SIFT Algorithm and KLT Matching Algorithm. *J. Astronaut.* **2011**, *7*, 028.
61. Xu, H.; Wang, Y.; Cheng, L.; Wang, Y.; Ma, X. Exploring a High-quality Outlying Feature Value Set for Noise-Resilient Outlier Detection in Categorical Data. In Proceedings of the Conference on Information and Knowledge Management (CIKM), Turin, Italy, 22–26 October 2018; pp. 17–26.
62. Sohail, A.S.M.; Bhattacharya, P. Classification of facial expressions using k-nearest neighbor classifier. In *Proceedings of the International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*; Springer: Berlin, Germany, 2007; pp. 555–566.
63. Wang, X.H.; Liu, A.; Zhang, S.Q. New facial expression recognition based on FSVM and KNN. *Optik* **2015**, *126*, 3132–3134. [CrossRef]
64. Valstar, M.; Patras, I.; Pantic, M. Facial action unit recognition using temporal templates. In Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication, Kurashiki, Okayama, Japan, 22 September 2004; pp. 253–258.
65. Michel, P.; El Kaliouby, R. Real time facial expression recognition in video using support vector machines. In Proceedings of the 5th International Conference on Multimodal Interfaces, Vancouver, BC, Canada, 5–7 November 2003; pp. 258–264.
66. Tsai, H.H.; Chang, Y.C. Facial expression recognition using a combination of multiple facial features and support vector machine. *Soft Comput.* **2018**, *22*, 4389–4405. [CrossRef]
67. Hsieh, C.C.; Hsih, M.H.; Jiang, M.K.; Cheng, Y.M.; Liang, E.H. Effective semantic features for facial expressions recognition using SVM. *Multimed. Tools Appl.* **2016**, *75*, 6663–6682. [CrossRef]
68. Saeed, S.; Baber, J.; Bakhtyar, M.; Ullah, I.; Sheikh, N.; Dad, I.; Sanjrani, A.A. Empirical Evaluation of SVM for Facial Expression Recognition. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 670–673. [CrossRef]
69. Shah, J.H.; Sharif, M.; Yasmin, M.; Fernandes, S.L. Facial expressions classification and false label reduction using LDA and threefold SVM. *Pattern Recognit. Lett.* **2017**. [CrossRef]
70. Wang, Y.; Ai, H.; Wu, B.; Huang, C. Real time facial expression recognition with adaboost. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 26 August 2004; Volume 3, pp. 926–929.
71. Liew, C.F.; Yairi, T. Facial expression recognition and analysis: A comparison study of feature descriptors. *IPSJ Trans. Comput. Vis. Appl.* **2015**, *7*, 104–120. [CrossRef]
72. Gudipati, V.K.; Barman, O.R.; Gaffoor, M.; Abuzneid, A. Efficient facial expression recognition using adaboost and haar cascade classifiers. In Proceedings of the Annual Connecticut Conference on Industrial Electronics, Technology & Automation (CT-IETA), Bridgeport, CT, USA, 14–15 October 2016; pp. 1–4.
73. Zhang, S.; Hu, B.; Li, T.; Zheng, X. A Study on Emotion Recognition Based on Hierarchical Adaboost Multi-class Algorithm. In *Proceedings of the International Conference on Algorithms and Architectures for Parallel Processing*; Springer: Berlin, Germany, 2018; pp. 105–113.
74. Moghaddam, B.; Jebara, T.; Pentland, A. Bayesian face recognition. *Pattern Recognit.* **2000**, *33*, 1771–1782. [CrossRef]
75. Mao, Q.; Rao, Q.; Yu, Y.; Dong, M. Hierarchical Bayesian theme models for multipose facial expression recognition. *IEEE Trans. Multimed.* **2017**, *19*, 861–873. [CrossRef]
76. Surace, L.; Patacchiola, M.; Battini Sönmez, E.; Spataro, W.; Cangelosi, A. Emotion recognition in the wild using deep neural networks and Bayesian classifiers. In Proceedings of the 19th ACM International Conference on Multimodal Interaction, Glasgow, UK, 13–17 November 2017; pp. 593–597.
77. Huang, M.W.; Wang, Z.W.; Ying, Z.L. A new method for facial expression recognition based on sparse representation plus LBP. In Proceedings of the 3rd International Congress on Image and Signal Processing, Yantai, China, 16–18 October 2010; Volume 4, pp. 1750–1754.
78. Zhang, S.; Zhao, X.; Lei, B. Facial expression recognition using sparse representation. *WSEAS Trans. Syst.* **2012**, *11*, 440–452.
79. El Emary, I.M.; Ramakrishnan, S. On the application of various probabilistic neural networks in solving different pattern classification problems. *World Appl. Sci. J.* **2008**, *4*, 772–780.

80. Kusy, M.; Zajdel, R. Application of reinforcement learning algorithms for the adaptive computation of the smoothing parameter for probabilistic neural network. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *26*, 2163–2175. [CrossRef]
81. Neggaz, N.; Besnassi, M.; Benyettou, A. Application of improved AAM and probabilistic neural network to facial expression recognition. *J. Appl. Sci. (Faisalabad)* **2010**, *10*, 1572–1579. [CrossRef]
82. Fazli, S.; Afrouzian, R.; Seyedarabi, H. High-performance facial expression recognition using Gabor filter and probabilistic neural network. In Proceedings of the IEEE International Conference on Intelligent Computing and Intelligent Systems, Shanghai, China, 20–22 November 2009; Volume 4, pp. 93–96.
83. Walecki, R.; Rudovic, O.; Pavlovic, V.; Schuller, B.; Pantic, M. Deep structured learning for facial expression intensity estimation. *Image Vis. Comput* **2017**, *259*, 143–154.
84. Breuer, R.; Kimmel, R. A deep learning perspective on the origin of facial expressions. *arXiv* **2017**, arXiv:1705.01842.
85. Liu, M.; Li, S.; Shan, S.; Wang, R.; Chen, X. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian Conference on Computer Vision*; Springer: Berlin, Germany, 2014; pp. 143–157.
86. Jung, H.; Lee, S.; Yim, J.; Park, S.; Kim, J. Joint fine-tuning in deep neural networks for facial expression recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2983–2991.
87. Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going deeper in facial expression recognition using deep neural networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10.
88. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion Aware Facial Expression Recognition Using CNN with Attention Mechanism. *IEEE Trans. Image Process.* **2019**, *28*, 2439–2450. [CrossRef]
89. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef]
90. Hinton, G.E.; Sejnowski, T.J. Learning and relearning in Boltzmann machines. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*; MIT Press: Cambridge, MA, USA, 1986.
91. Liu, P.; Han, S.; Meng, Z.; Tong, Y. Facial expression recognition via a boosted deep belief network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 1805–1812.
92. Zhao, X.; Shi, X.; Zhang, S. Facial expression recognition via deep learning. *IETE Tech. Rev.* **2015**, *32*, 347–355. [CrossRef]
93. He, J.; Cai, J.; Fang, L.; He, Z.; Amp, D.E. Facial expression recognition based on LBP/VAR and DBN model. *Appl. Res. Comput.* **2016**.
94. Uddin, M.Z.; Hassan, M.M.; Almogren, A.; Alamri, A.; Alrubaian, M.; Fortino, G. Facial expression recognition utilizing local direction-based robust features and deep belief network. *IEEE Access* **2017**, *5*, 4525–4536. [CrossRef]
95. Wöllmer, M.; Kaiser, M.; Eyben, F.; Schuller, B.; Rigoll, G. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image Vis. Comput.* **2013**, *31*, 153–163. [CrossRef]
96. Kim, D.H.; Baddar, W.; Jang, J.; Ro, Y.M. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Trans. Affect. Comput.* **2017**. [CrossRef]
97. Hasani, B.; Mahoor, M.H. Facial expression recognition using enhanced deep 3D convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 2278–2288.
98. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canad, 8–13 December 2014; pp. 2672–2680.
99. Lai, Y.H.; Lai, S.H. Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition. In Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 263–270.

100. Zhang, F.; Zhang, T.; Mao, Q.; Xu, C. Joint pose and expression modeling for facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 19–21 June 2018; pp. 3359–3368.
101. Yang, H.; Zhang, Z.; Yin, L. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks. In Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 294–301.
102. Chen, J.; Konrad, J.; Ishwar, P. Vgan-based image representation learning for privacy-preserving facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1570–1579.
103. Yang, H.; Ciftci, U.; Yin, L. Facial expression recognition by de-expression residue learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18 June 2018; pp. 2168–2177.
104. Lundqvist, D.; Flykt, A.; Öhman, A. The Karolinska directed emotional faces (KDEF). *CD ROM Dep. Clin. Neurosci. Psychol. Sect. Karolinska Inst.* **1998**, *91*, 630.
105. Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. Web-based database for facial expression analysis. In Proceedings of the IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005; p. 5.
106. Kaulard, K.; Cunningham, D.W.; Bülthoff, H.H.; Wallraven, C. The MPI facial expression database—A validated database of emotional and conversational facial expressions. *PLoS ONE* **2012**, *7*, e32321. [CrossRef]
107. Lucey, P.; Cohn, J.F.; Prkachin, K.M.; Solomon, P.E.; Matthews, I. Painful data: The UNBC-McMaster shoulder pain expression archive database. In Proceedings of the Face and Gesture 2011, Santa Barbara, CA, USA, 21–25 March 2011; pp. 57–64.
108. Mavadati, S.M.; Mahoor, M.H.; Bartlett, K.; Trinh, P.; Cohn, J.F. Disfa: A spontaneous facial action intensity database. *IEEE Trans. Affect. Comput.* **2013**, *4*, 151–160. [CrossRef]
109. Yin, L.; Wei, X.; Sun, Y.; Wang, J.; Rosato, M.J. A 3D facial expression database for facial behavior research. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 10–12 April 2006; pp. 211–216.
110. Zhang, X.; Yin, L.; Cohn, J.F.; Canavan, S.; Reale, M.; Horowitz, A.; Liu, P.; Girard, J.M. Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database. *Image Vis. Comput.* **2014**, *32*, 692–706. [CrossRef]
111. Wang, S.; Liu, Z.; Lv, S.; Lv, Y.; Wu, G.; Peng, P.; Chen, F.; Wang, X. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Trans. Multimed.* **2010**, *12*, 682–691. [CrossRef]
112. Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; Baker, S. Multi-PIE. *Image Vis. Comput.* **2010**, *28*, 807–813. [CrossRef]
113. Zhao, G.; Huang, X.; Taini, M.; Li, S.Z.; Pietikäinen, M. Facial expression recognition from near-infrared videos. *Image Vis. Comput.* **2011**, *29*, 607–619. [CrossRef]
114. Carrier, P.L.; Courville, A.; Goodfellow, I.J.; Mirza, M.; Bengio, Y. *FER-2013 Face Database*; Universit de Montral: Montreal, QC, Canada, 2013.
115. Valstar, M.F.; Mehu, M.; Jiang, B.; Pantic, M.; Scherer, K. Meta-analysis of the first facial expression recognition challenge. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2012**, *42*, 966–979. [CrossRef]
116. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimed.* **2012**, *19*, 34–41. [CrossRef]
117. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 2106–2112.
118. Li, S.; Deng, W.; Du, J. 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), Honolulu, HI, USA, 22–25 July 2017; pp. 2852–2861.
119. Li, S.; Deng, W. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Trans. Image Process.* **2019**, *28*, 356–370. [CrossRef]
120. Li, S.; Deng, W. Blended Emotion in-the-Wild: Multi-label Facial Expression Recognition Using Crowdsourced Annotations and Deep Locality Feature Learning. *Int. J. Comput. Vis.* **2018**, 1–23. [CrossRef]

121. Whitehill, J.; Littlewort, G.; Fasel, I.; Bartlett, M.; Movellan, J. Toward practical smile detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 2106–2111. [CrossRef]
122. Jeon, J.; Park, J.C.; Jo, Y.; Nam, C.; Bae, K.H.; Hwang, Y.; Kim, D.S. A Real-time Facial Expression Recognizer using Deep Neural Network. In Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, Danang, Vietnam, 4–6 January 2016; p. 94.
123. Hossin, M.; Sulaiman, M. A review on evaluation metrics for data classification evaluations. *Int. J. Data Min. Knowl. Manag. Process* **2015**, *5*, 1.
124. Li, S.; Deng, W. Deep Emotion Transfer Network for Cross-database Facial Expression Recognition. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 3092–3099.
125. Li, S.; Deng, W. A Deeper Look at Facial Expression Dataset Bias. *arXiv* **2019**, arXiv:1904.11150.
126. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.* **2017**. [CrossRef]
127. Ramakrishnan, S.; El Emary, I.M. Speech emotion recognition approaches in human computer interaction. *Telecommun. Syst.* **2013**, *52*, 1467–1478. [CrossRef]
128. Chang, J.; Scherer, S. Learning representations of emotional speech with deep convolutional generative adversarial networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2746–2750.
129. Chao, L.; Tao, J.; Yang, M.; Li, Y.; Wen, Z. Multi-scale temporal modeling for dimensional emotion recognition in video. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Orlando, FL, USA, 7 November 2014; pp. 11–18.
130. Chen, S.; Jin, Q. Multi-modal dimensional emotion recognition using recurrent neural networks. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge Brisbane, Australia, 26 October 2015; pp. 49–56.
131. He, L.; Jiang, D.; Yang, L.; Pei, E.; Wu, P.; Sahli, H. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, Brisbane, Australia, 26 October 2015; pp. 73–80.
132. Newton, E.M.; Sweeney, L.; Malin, B. Preserving privacy by de-identifying face images. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 232–243. [CrossRef]
133. Rahulamathavan, Y.; Rajarajan, M. Efficient privacy-preserving facial expression classification. *IEEE Trans. Dependable Secur. Comput.* **2017**, *14*, 326–338. [CrossRef]