IJMIC Vol. 1, Issue 2, 1-25

Review Article

Static Facial Expression Recognition in the Wild: Taxonomy, Trends and Challenges

Jing-Zhi Koay[1], Jason Teo[1,2,3*]

[1]Faculty of Computing and Informatics, Jalan UMS, 88400 Kota Kinabalu, Sabah, Universiti Malaysia Sabah
[2]Creative Advanced Machine Intelligence Research Centre, Faculty of Computing and Informatics, Jalan UMS, 88400 Kota Kinabalu, Sabah, Universiti Malaysia Sabah
[3]Evolutionary Computing Laboratory, Faculty of Computing and Informatics, Jalan UMS, 88400 Kota Kinabalu, Sabah, Universiti Malaysia Sabah

*Corresponding author: Jason Teo, jtwteo@ums.edu.my

Received: [March 6, 2025]
Accepted: [April 29, 2025]
Published: [July 31, 2025]

Abstract - In recent years, Facial Expression Recognition (FER) has gained significant attention due to its wide application and potential in various domains. FER is the research field that focuses on recognizing and classifying human emotions expressed by humans into emotion categories using computer vision. Different machine learning techniques have been applied to this research field with promising outcomes through the application of increasingly more powerful machine learning algorithms. This systematic literature review is conducted to investigate static FER on unconstrained datasets. A total of 32 studies were retrieved from four major academic repositories. The aim of this study is to provide a comprehensive review of static FER research on unconstrained facial expression image datasets including the overview of key concepts, the approaches applied, the datasets used, the current state-of-the-art as well as the future directions of research in this fast-developing research field. Deep learning methods emerged as the most promising approach for static FER while second-order pooling in CNNs allowed for improved representation of regional features and facial landmark distortion.

Keywords: emotion recognition, facial recognition, machine learning, computer vision, affective computing

1. INTRODUCTION

Facial expression is one of the most powerful, natural, and universal signals for human beings to convey their emotion and intentions. Facial expression refers to the movement of facial muscles that reveals the emotions of a person. Among several nonverbal communication such as gestures, voice tone, and body language, facial expressions are one of the most important and meaningful ways of human communication. Facial expression is arguably the most important clue for analysing one's genuine emotions. Ekman and Friesen [1] defined six basic expressions, including anger, disgust, fear, happiness, sadness, and surprise are considered universal emotional expressions among all people.

Facial Expression Recognition (FER) has attracted great interest from industry and research because of its huge potential in various domains such as education, healthcare, and entertainment. Many real-world applications such as clinical human emotion analysis, affective-aware video games, and e-tutoring platforms exploit facial expressions as an input that conceals human emotion. FER is also particularly useful in market research and consumer testing, virtual assistants and smart devices, automotive safety systems, mental health monitoring apps, security screening and pain assessment tools to name but a few. Many researchers have made significant progress in the field of FER by applying different machine learning techniques.

The aim of FER is to identify human facial expressions from the camera sensor into emotion categories by detecting patterns from their facial features. However, facial expression may vary from one individual to another due to variations in age, gender, pose, angles, illumination, and the presence of occlusions. Therefore, this study will focus on FER studies on uncontrolled facial expression image datasets which truly reflect the natural facial expressions in real-life.

Generally, FER consists of three main stages: face detection and pre-processing, facial feature extraction, and classification of expressions [2, 3, 4]. First, a face detector is used to spot the face from the scene, then facial landmarks such as eyes and nose are located from the face. Next, the captured face will undergo feature extraction to produce a representation of face data that better describe the expression. After that, the extracted features are fed into the classifier to perform expression recognition. The details of each stage will be further explained in Section 4.

This study presents a systematic review of static Facial Expression Recognition (FER) on the spontaneous facial expression dataset by reviewing existing literature in this specific research domain. The purpose is to illustrate an overview of research works, including the key concepts, the approaches, the dataset, and the current state-of-the-art as well as the future directions in this field. The contributions of this article are as follows:

1. Summarize the main tasks and classify the existing methods for each of these tasks.
2. Update the current state of research in this field.
3. Identify the most used dataset.
4. Describe the techniques and the most widely adapted configurations.
5. Identify the most used performance evaluation protocol.
6. Highlight the future directions for research in this field.

The rest of this systematic review is organized as follows: Section 2 presents the background along with the main task in FER; Section 3 presents the paper selection criteria for this systematic review, including a quantitative analysis of the papers gathered; Section 4 presents an analysis of the gathered papers to answer the research questions by describing the methods employed for FER as well as their popularity, the strengths and drawbacks of each method, and the main challenges in this field; Section 5 covers the evaluation and findings from the considered works; Section 6 presented the conclusions.

2. BACKGROUND

2.1 Face detection and alignment

Generally, the first step in FER is detecting the face in the image. Face detection is responsible for locating the region of interest (ROI) of the face image and hence subtracting the background

for excluding irrelevant features in the image. Since the detected face in an input image does not necessarily in a proper condition for further analysed. Face alignment is required to reduce the variation in facial size and in-plane rotation. In short, face alignment is a rotation operation that aligns the landmark position of the eyes to be parallel with the horizontal axis.

2.2 Pre-processing

Data pre-processing is the key operation in any machine learning algorithm, including the field of FER. Variations that are not related to a facial expression such as illuminations, backgrounds, and head pose commonly exist in unconstrained datasets. Issues in the unconstrained dataset such as irrelevant image-to-facial expressions, uncontrolled illuminations, and highly variable posture could greatly depreciate the accuracy. For example, the variation of illumination in the input image data may hinder the feature extraction and hence affect the accuracy of FER. Therefore, it is necessary to perform pre-processing on the dataset to obtain better performance. Besides that, the data augmentation technique is often used to increase the diversity of the images to prevent over-fitting problems, especially for the deep learning approaches which require a huge amount of training data.

2.3 Feature Extraction

Feature extraction is conducted to extract relevant representations or descriptions of the input image. The quality of extracted features is the key in guaranteeing efficient and accurate expression recognition. The early works in FER mostly rely on handcrafted features by using manual feature extraction. Generally, feature extraction techniques are divided into Geometric-based and Appearance-based, Action Unit (AU) based and Non-AU based, Local and Holistic [5]. Geometric-based techniques such as Optical Flow, and Active Appearance Model, are based on extracting shapes and locations of facial landmark points including mouth, eyes, and nose to form a feature vector that represents the face geometry. In contrast, appearance-based techniques simply apply an image filter such as Gabor wavelets to extract feature vectors [4]. Examples of appearance-based methods for handcrafted features are Histogram of Oriented Gradients (HOG), Local Binary Pattern (LBP), Local Phase Quantization (LPQ), and Gabor Filter. AU are descriptions of facial movements based on the face muscle contraction. The facial action coding system (FACS) [1] employed a set of Action Units (AUs) to describe the facial actions and their intensity. Holistic approaches take the face as a whole for feature processing, while local approaches focus on the facial regions that are prone to change with facial expressions [6]. Wang et al. [7] presented a comprehensive review of facial feature extraction including the challenges in the wild environment. Some research applied Principal Component Analysis (PCA) [8] in feature selection to reduce the dimensionality of features with high variance [9]. The PCA algorithm can effectively remove redundant features while preserving details, resulting in an increase in computational efficiency [10].

Apart from handcrafted features, automatic feature extraction can be achieved through deep learning-based methods to extract appearance-based feature representation. The automatic feature extraction method is widely adopted in the field of FER with the emergence of Deep Learning and outperforms the manual extraction method in most cases. In contrast to handcrafted features, deep features are extracted directly from the input data through the convolution layer and pooling layer with iterative algorithms like gradient descent [11]. The convolution layer is responsible to extract elementary features such as edges, corners, and shapes from the input images or the feature maps from the previous layer, while the pooling layer reduces the dimensionality of the feature maps by applying a function such as max pooling or average pooling [12].

2.4 Classification

In this stage, the extracted features are fed into a classifier to identify the categories of facial expressions (anger, disgust, fear, happiness, sadness, surprise, and neutral). Generally, the classifiers can be divided into conventional Machine Learning (ML) based classifiers including Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Neural Network (NN) or Deep Learning (DL) based classifiers such as Convolutional Neural Network (CNN). Both methods were used in the reviewed papers, CNN being the most popular choice while only a few papers used SVM for classification. CNN is a type of neural network that is mainly used to deal with pattern recognition and object detection problem [13]. The reason CNN is considered above all other classification algorithms is the ability to automatically discover the multiple levels of representations in data with higher levels representing more abstract concepts [14]. Those features are then used for classification, generally through a Softmax function that retrieves the highest probability from the classes' probability distribution as the predicted class. Tang [15] presented a framework that replaced the last softmax layer with a linear SVM. The proposed framework outperforms all other methods in the FER challenge of the ICML 2013 Workshop hosted on Kaggle [16].

Table 1: Summary of FER methods

Approach	Key Strengths	Key Limitations
Conventional ML (HOG, LBP, SVM, KNN)	Better interpretability, less computationally intensive, works with smaller datasets	Manual feature extraction, lower accuracy on complex datasets, less effective for non-linear patterns
Deep Learning (CNN-based models)	Automatic feature extraction, end-to-end training, state-of-the-art accuracy, better pattern learning	Limited interpretability, requires large datasets, computationally expensive, overfitting risks
Hybrid approaches (CNN+SVM)	Combines strengths of both approaches, avoids manual feature extraction, improved interpretability	More complex methodology, additional implementation complexity
Ensemble methods	Best overall performance, improved robustness, particularly effective for diverse datasets (e.g., FER-2013)	Higher computational cost, careful classifier selection needed, complex implementation
Enhancement Techniques		
Covariance Pooling	Better regional feature representation, 2-4% higher accuracy on RAF-DB, 5-10% on SFEW	Increased computational complexity
Face Detection & Alignment	MTCNN combines detection and alignment, reduces pose variation effects	Viola-Jones degrades with non-frontal faces, variable benefits across datasets
Data Augmentation	Addresses overfitting with limited data, effective for SFEW and RAF-DB	Can decrease accuracy if excessive, less effective for large datasets
Attention Mechanisms	Improves feature extraction, focuses on emotion-relevant areas	Adds model complexity, variable generalization across expressions

Table 1 summarizes the main strengths and weaknesses of the various FER approaches that were reviewed in this study.

Figure 1 shows a high-level overview of the taxonomy of FER methods.

3. METHODOLOGY

This section describes the PRISMA methodology adopted to accomplish this systematic review and the process of retrieving, filtering, analysing, and extracting the existing literature of FER

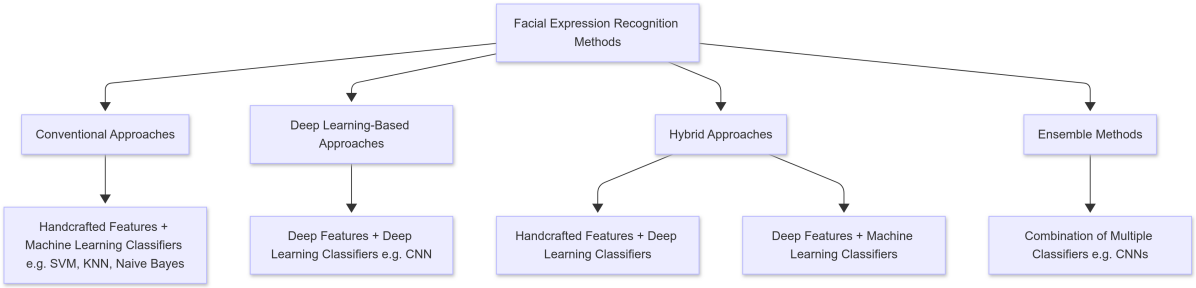

Figure 1: Taxonomy of FER methods

research on spontaneous facial expression image datasets. This section consists of four subsections: the first presents the research question; the second on the source of data; the third on the selection criteria; the fourth on the analysing of considered studies.

3.1 Research Questions

This systematic review is conducted to answer research questions as listed below:

1. What are the processes in image-based Facial Expression Recognition of spontaneous unconstrained facial expression dataset?
2. Which methods are widely used for conducting research in this field?
3. What are the strengths and drawbacks of the method?
4. What is the widely adapted dataset in this field?
5. What are the main challenges within this field?

3.2 Source of Data

In order to conduct a systematic review, it is expedient to select rich and standard databases as the source of data. The following digital libraries and databases were accessed: ACM Digital Libray (https://dl.acm.org), IEEE Xplore (https://ieeexplore.ieee.org), Springer Link (https://link.springer.com); and ScienceDirect (https://www.sciencedirect.com). These repositories were selected for data retrieval because of the availability of high-impact journals and conferences. These repositories were queried with the following search string: (”facial expression recognition”) AND (”static” OR ”image-based” OR ”image-based”) AND (”unconstrained” OR ”spontaneous” OR ”wild”). As a result, a total number of 273 papers were retrieved as shown in Table 2 below.

Table 2: Number of papers by repository

Repository	ACM Digital Library	IEEE Xplore	ScienceDirect	Springer Link
Number of papers	59	8	100	106

3.2.1 Selection Criteria

Due to the large number of papers that have been retrieved, several selection criteria were established to remove irrelevant and redundant papers. The refinement was performed based on the criteria listed below:

- Restricting the search to journals and conference papers.
- Selecting computer science and IT-related papers.

By eliminating papers that are not related and duplicated, a total number of 83 seemingly relevant papers were left. A manual review of the remaining papers was performed to exclude studies that do not contribute to answering the research questions. This stage was conducted by reading through the abstracts and content in case the abstracts were not clear enough to determine whether they facilitate this systematic review. Additionally, the bibliography cited in the papers that passed this stage was also reviewed. As a result, the final set of retrieved data consists of 32 papers that are useful and appropriate to answer the research questions, the process as shown in Figure 2.

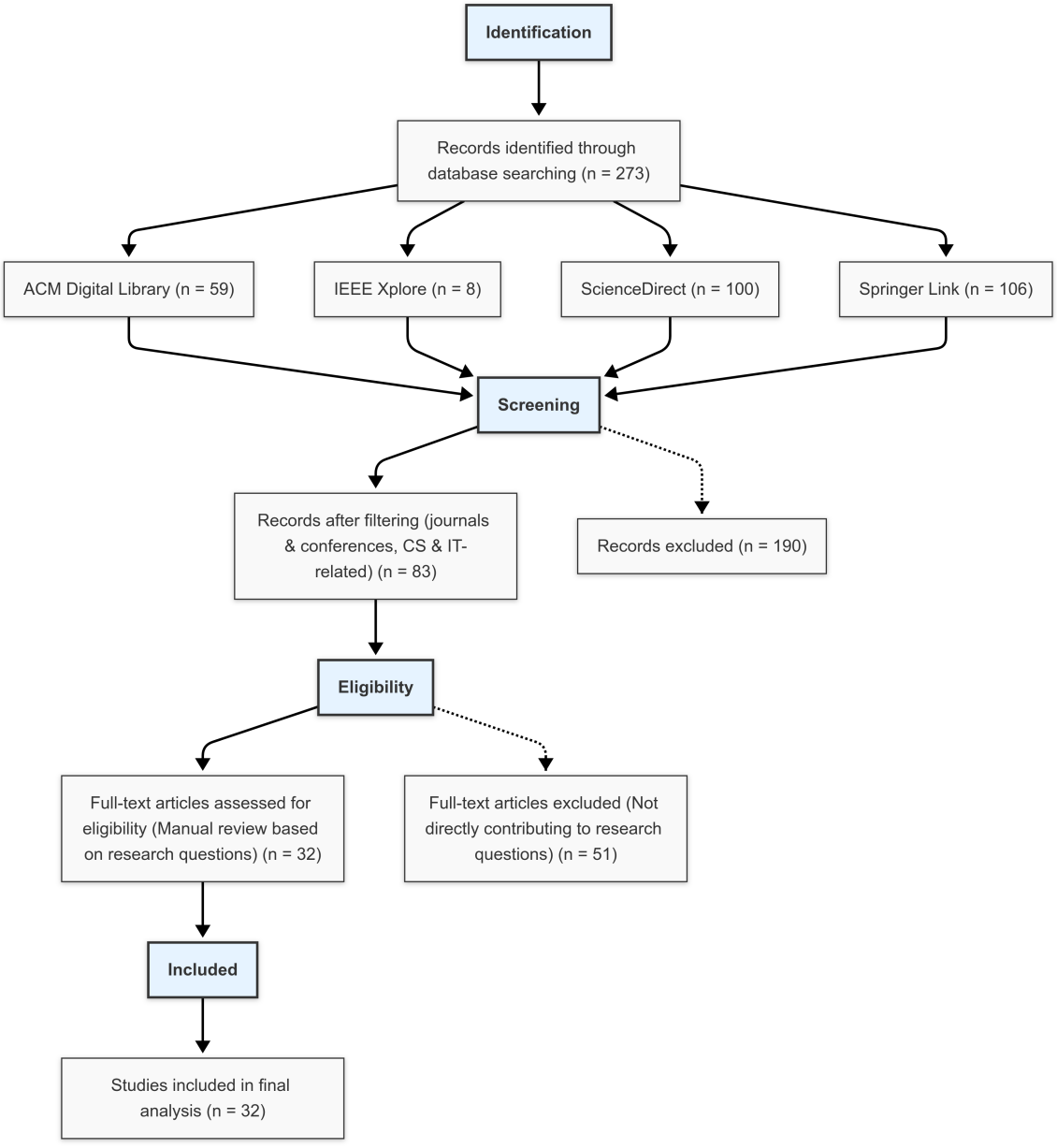

Figure 2: Flowchart of the PRISMA process

3.3 Summary of Final Papers Selected for Analysis

This sub-section presents an overview summary of the final retrieved papers to be analyzed for the systematic review.

Table 3 summarizes the number of papers published and their publication venue. Twenty-three

Table 3: Number of papers by publication type

Type	Publication	Number
Conference	ACM International Conference on Multimedia in Asia (MM'Asia)	1
	IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop	1
	International Conference on Pattern Recognition (ICPR)	1
	International Symposium on Signal, Image, Video and Communications (ISIVC)	1
	ACM International Conference on Multimedia (MM)	2
	IEEE International Conference on Image Processing (ICIP)	1
	International Workshop on the Future of Internet of Everything (FIoE)	1
	Internation Workshop on Human-centric Multimedia Analysis (HuMA)	1
	Total	9
Journal	Cluster Computing	1
	International Journal of Computer Vision	1
	Multimedia Tools and Application	4
	Neurocomputing	6
	Journal of Visual Communication and Image Representation	2
	Optik	1
	Digital Object Identifier	1
	Expert Systems with Applications	1
	Image and Vision Computing	1
	Pattern Recognition	1
	IEEE Transactions on Neural Networks and Learning Systems	1
	Cognitive Systems Research	1
	The Visual Computer	2
	Total	23

out of thirty-two of them have been published in journals (71.88%). The *Journal of Neurocomputing* from ScienceDirect has the most number of publications at 6, followed by the *Journal of Multimedia Tools and Application* journal from Springer which has 4 publications, then both the *Journal of The Visual Computer* from Springer and *Journal of Visual Communication and Image Representation* from ScienceDirect has 2 publications for each. The *ACM International Conference on Multimedia* also contributed 3 publications to the collection.

4. RESULTS & DISCUSSIONS

4.1 Face Detection and Pre-processing

4.1.1 Face Detection and Alignment

For the first stage of face detection and alignment, several face detection methods are commonly used including the Viola-Jones face detector aka Haar Cascade Classifier [17], Multitask Cascade Convolutional Neural Network (MTCNN) [18], and Faster RCNN [19]. Certain research applies both face and facial landmark detection in order to perform face alignment, facial landmark detectors such as One-millisecond algorithm [20], Discriminative Response Map Fitting [21] and

MTCNN are used in the retrieved collections of literature. Besides that, there is also research that employed predefined Convolutional Neural Networks (CNN) namely ResNet-10 [22] as face detector [23] and the Lib face detection algorithms that support 68 landmarks detection [24].

According to the thirty-two reviewed papers, the Viola-Jones face detector is the most used face detector [25, 26, 27, 28, 29, 30, 31, 32], followed by MTCNN [33, 34, 35, 36, 37, 38], while the rest are less preferred. For face alignment, MTCNN [33, 36, 38, 35] and One milliseconds algorithm [39, 40] are more popular than Lib face detection algorithm [41]. Some research claimed that merging face detection and facial landmark detection for face alignment such as MTCNN yields better results. However, some research does not perform face alignment for particular databases such as RAF-DB [42] and SFEW [43] where the images are already aligned, and FER-2013 [16] because the faces are approximately centered and occupied the same area in each image. Additionally, there is also research that do not use face detection nor face alignment [44, 45, 46, 47, 48, 49, 50].

4.1.2 Pre-processing

Based on the reviewed papers, several pre-processing techniques include face cropping, image resizing, grayscale conversion, normalization, histogram equalization, and Fast Fourier Transform. Face cropping and image resizing are often used together to enlarge the facial details and to fit the input size of CNN, especially in research that involved multiple datasets with different image sizes. Additionally, grayscale conversion is useful for ensuring consistency between the colored image datasets and grayscale image datasets. Normalization techniques such as Mix-Max Normalization, Spatial Normalization, Scale Normalization, and Mean Normalization are categorized as geometric normalization that helps to rescale the pixel value into a range without distorting differences in the values. Histogram equalization is useful to enhance the details for better feature extraction by increasing the global contrast and also aids the largely varying illumination in the unconstrained datasets. Beside Histogram Equalization, there are also other illumination normalization techniques such as Isotropic Diffusion-based Normalization (Isotropic Smoothing) and Local Binary Pattern Histogram-based Normalization (LBP Histogram) [51] which are effective in reducing the effect of varying illumination which is common in the unconstrained dataset where the image data are collected in an uncontrolled environment.

It is worth noting that almost one-third of the reviewed papers have applied data augmentation to improve the generalization ability of the Deep Learning models and hence increase the recognition performance. Data augmentation is a technique that generates additional training data by applying random perturbations such as rotation, shifts, skew, scaling, and flips on the original dataset [52]. Data augmentation helps to alleviate the over-fitting problem of the model due to insufficient training data by increasing the diversity of training data. Additionally, data augmentation also can effectively enhance the robustness of the deep learning classifiers. Most prior works use a standard data augmentation method that is not specific to the FER task.

According to the reviewed papers, researchers commonly apply multiple pre-processing techniques rather than only one. For example, Renda et al. [32] applied Histogram Equalization, Isotropic Smoothing, image resizing, grayscale conversion, random cropping, and data augmentation. Despite the trend, some researchers do not apply any pre-processing techniques [29, 44, 34, 49, 37, 50, 53] which is seven out of thirty-two papers (21.88%).

4.2 Feature Extraction and Classification

According to the reviewed papers, methods for extracting handcrafted features include Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), Local Phase Quantization (LPQ), Dual-Cross Patterns (DCP), Local Binary Features (LBF), PCA, Speeded-up Robust

Features (SURF) descriptor [54], and novel method such as Multistage Binary Patterns (MSBP) [26] and Merged Binary Patterns Code (MBPC) [55]. Besides handcrafted features, there are also deep features that can be extracted automatically using Deep Learning models namely CNN.

Generally, the FER approaches can be divided into two groups depending on the feature extraction method. Conventional approaches fed handcrafted features into ML classifiers such as SVM and KNN, while DL-based approach such as CNN does not require manual feature extraction as it was able to perform end-to-end FER, from feature extraction to expression classification. In contrast, conventional approaches rely on manually extracted features which is laborious and time-consuming whereas DL-based approaches conduct both feature extraction and classification in a unified framework. Additionally, DL-based methods can learn more hidden data representations through multiple levels of abstraction than conventional ML-based methods [56].

Surprisingly, despite the DL-based approach showing several advantages over conventional approaches, it does not significantly dominate in the reviewed papers, with only half of them being DL-based. The other half of reviewed papers falls on conventional approaches such as SVM, KNN, Neural Network, Naive Bayes, and Decision Tree. Instead of using solely deep features on the DL classifiers, some researchers fed handcrafted features to DL classifiers or use ML classifiers on deep features.

Unsurprisingly, DL-based methods were popular with half of the reviewed papers employing DL-based approaches. The possible reasons are DL method able to conduct end-to-end training to classification and less depending on pre-processing. The conventional approaches come next in popularity with seven out of thirty-two at (21.88%), followed by the combination of deep features and ML classifier at (15.62%). For example, the work in [57, 27] applied SVM on deep features from a predefined DL model and the work in [34, 35] used an ensemble model of CNN for feature extraction and SVM for classification.

Furthermore, ensemble methods such as bagging and boosting are applied to combine multiple classifiers for better accuracy and robustness. Two studies had employed ensemble classifiers, by combining multiple CNN models. Zia et al. [29] presents a dynamically weighted majority-based architecture for an ensemble model, and Renda et al. [32] has examined multiple strategies in forming ensemble classifiers such as majority voting, aggregating, and averaging in application to FER.

Despite the benefit of ensemble methods, it does not guarantee yielding better accuracy than a single classifier, and it requires a good understanding of the problem and data when choosing ensemble strategies as there are many ways to form ensemble models. In addition, ensemble models are more costly to create and train due to their complexity and lack of interpretability.

For dimensionality reduction, PCA is used in three reviewed papers to reduce redundant features while preserving the discriminative features. However, it is noteworthy that no research that confirms the necessity of PCA in this field was found in the reviewing of retrieved papers.

4.3 Datasets

The prerequisite of developing any FER algorithm is having adequate labeled facial expression data. Generally, the facial expression datasets can be divided into two types depending on the method to collect samples, one being posed datasets while another type is the unconstrained dataset. For posed facial expression dataset, the facial expression data are collected by actors who were asked to artificially generate particular expressions with frontal faces under controlled lighting conditions, hence it may not reflect the natural facial expressions in real-world situa-

tions. Therefore, it may be the bottleneck for FER application in the real world as it ignores the real world such as varying illumination, image quality, presence of occlusions, background noise, and head pose. In contrast, the unconstrained dataset which is also known as the wild dataset contains spontaneous facial expressions from a highly uncontrolled environment. The facial expressions are rather naturally evoked and consistent with the underlying emotional state, often elicited by capturing participants' reactions from watching emotional stimuli videos or playing video games.

The image acquisition protocols for non-posed facial expression data have been discussed in Saha et al. [58]. This author highlighted a few important factors to be considered in facial expression acquisition: Deliberate or emotion-elicited facial expressions? Who expresses emotions? Controlled or uncontrolled environment? Participant's awareness of the experiment? Furthermore, the author claimed that emotional video clips are the best at inducing spontaneous facial expressions, and most of the unconstrained facial expression datasets are designed in this way.

Based on the reviewed papers, the FER-2013 [16], SFEW [43], RAF-DB [42], MMI database [59], FEEDTUM database [60], BAUM-2i [61], AffectNet [62] are used. Among all the datasets, SFEW, FER-2013, RAF-DB are considered the most widely used unconstrained datasets in the prior study. All the datasets included in this review are considered to be the standard datasets commonly used in FER comparison studies.

FER-2013 is an online open dataset for facial expressions consisting of more than 30,000 face images, including 28,709 images for the training set, and 3589 images for each validation and test set. All the images are grayscale images of size 48x48 pixels and labeled with six basic facial expressions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The image in FER-2013 was taken from the internet by using Google image search API with a set of emotion-related keywords and a combination of words related to gender, age, or ethnicity.

SFEW was created by selecting static frames from the AFEW dataset that consists of video clips collected from different movies with vast variations. It has over 1,700 colored facial images assigned to one of the six basic expressions and neutral. However, 372 images categorized as test sets are not publicly available for challenging purposes. Therefore, most research only utilized the training and validation sets of SFEW for the training and test phases. The training set contains 847 images and the validation set contains 409 images.

RAF-DB is a large-scale facial expression dataset that includes almost 30,000 uncontrolled images downloaded from the Internet. All the images are RGB images in 100x100 pixels. Unlike other datasets where each image is assigned with one label, this dataset has two subsets, the first single-label subset contains only one label out of the six basic expressions and neutral, while the second subset contains a multi-label facial image with up to 11 class compound emotions. Additionally, RAF-DB also provides face bounding boxes, manually and automatically annotated facial landmarks, and aligned facial images in the dataset. All research in the reviewed papers used only a single-label subset that consists of 12,271 images for training and 3068 images for testing.

Table 4 summarizes the key highlights from each of the three main datasets reviewed above.

4.4 Discussion

The review shows most works that produced top-tier results are achieved by DL-based approaches regardless of the dataset. DL-based approaches have dominated all reviewed works that research on FER-2013 dataset as shown in Table 5. For example, the highest accuracy for the FER-2013 test set was 72.25%, achieved by Renda et al. [32] using an ensemble of 9

Table 4: Summary of datasets

Dataset	Image Type	Resolution	Classes	Key Characteristics
FER-2013	Grayscale	48 × 48 pixels	7 (six basic + neutral)	Internet-sourced via Google API approximately centered faces
SFEW	Color	Variable	7 (six basic + neutral)	Static frames from movie clips, varying illumination and poses
RAF-DB	Color	100 × 100 pixels	7 (six basic + neutral for single-label subset) and 11 for multi-label subset)	Internet-sourced; includes face landmarks and alignment

pre-trained CNNs. Shao and Qian [37] shows next to best result at 71.14% using a pre-trained CNN namely ResNet101. For the SFEW dataset as shown in Table 6, the highest accuracy on the validation set is 59.86% [23] using a novel Deep Disturbance-disentangled Learning Model pre-trained on other datasets. Tong et al. [36] shows a comparable result at 59.52% without pre-training on other datasets by using a DL-based approach, which is a CNN called DenseNet with covariance pooling, followed by 58.14% from Acharya et al. [33] using Deep Locality-Preserving (DLP) CNN with covariance pooling too. Based on Table 7, Tong et al. [36] also recorded the highest accuracy on the RAF-DB test set at 88.63% using the same method, while the second highest accuracy was 87.71% by the novel method from Ruan et al. [23], then Acharya et al. [33] scored the third highest accuracy at 87.00% using the DLP-CNN. The top 3 results for SFEW and RAF-DB are from the same work [23, 36, 33], which shows their approaches are good choices for these two datasets and have a fair amount of robustness.

Table 5: Results of Reviewed Papers on FER-2013

Literature	Pre-processing	Feature Extraction	Classifier	Evaluation Protocol	Accuracy
Xie et al. [28]	Viola-Jones face detector, resize to 224x224	DAM-CNN (VGG-Face like CNN), SERD, MPVS-Net	DAM-CNN (VGG-Face like CNN), SERD, MPVS-Net	test set	66.20%
Renda et al. [32]	data augmentation and resize to 54x54 pixels	classical feed-forward CNN	Ensemble of 9 pre-trained CNNs	10 crop oversampling on test set	72.25%
Shao and Qian [37]	-	pre-trained CNN (ResNet101)	pre-trained CNN (ResNet101)	test set	71.14%
Fei et al. [27]	Viola-Jones face detector and face cropping	pre-trained CNN (AlexNet)	LDA	5-fold cross-validation	56.40%

Table 6: Results of Reviewed Papers on SFEW

Literature	**Pre-processing**	**Feature Extraction**	**Classifier**	**Evaluation Protocol**	**Accuracy**
Acharya et al. [33]	MTCNN face and facial landmark detection, face alignment, data augmentation	DLP-CNN with covariance pooling	DLP-CNN with covariance pooling	validation set	58.14%
Dapogny et al. [44]	-	HoG	novel local expression predictions (Random Forest)	10-fold cross-validation	37.10%
Luo et al. [57]	face alignment, image cropping and resize to 90x90 pixels, min-max normalization, data augmentation	VGG-like Deep CNN with local subclass constraint	SVM (RBF kernel)	validation set	55.03%
Zhang et al. [41]	face detection, resize to 256x256 pixels, manually cropping	TSNE and KNN	ResNet50	5-fold cross-validation	30.75%
Sun et al. [46]	face detection and alignment, isotropic diffusion	VGG-like Deep CNN with visual attention mapping	VGG-like Deep CNN with visual attention mapping	validation set	40.0%
Xie et al. [28]	Viola-Jones face detector, resize to 224x224	DAM-CNN (VGG-Face like CNN), SERD, MPVS-Net	DAM-CNN (VGG-Face like CNN), SERD, MPVS-Net	validation set	42.30%
Ji et al. [49]	-	ICID fusion network with DarkNet-19 CNN	DarkNet-19 CNN	validation set	51.2%
Tong et al. [36]	MTCNN face detector and data augmentation	DenseNet121 with covariance pooling	DenseNet121 with covariance pooling	validation set	59.52%

Sadeghi and Raie [63]	face detection, geometric and illumination normalization	novel Local Chi-squared Metric Learning algorithm	SVM	validation set	55.50%
Ruan et al. [23]	face detector, resize to 100x100 pixels, random cropping to 90x90 pixels, data augmentation	pre-trained novel Disturbance Feature Extraction Model (DFEM)	pre-trained novel Disturbance Feature Extraction Model (DFEM)	validation set	59.86%
Ly et al. [34]	MTCNN face detector	Ensemble CNNs (Inception-ResNet V1 and PointCNN)	SVM	validation set	56.20%
Ly et al. [35]	MTCNN face detector	Ensemble CNNs (Inception-ResNet V1 and PointCNN)	SVM	validation set	56.20%
Sadeghi and Raie [40]	one millisecond face and facial landmark detector, face cropping, normalization to 150x110 pixels, grayscale conversion	Gabor filtering	SVM (polynomial kernel)	validation set	36.92%
Wang et al. [64]	face detector (ResNet-10 from OpenCV), face landmark detector, data augmentation, random cropping and rescaling to 224x224, grayscale conversion	ResNet18 with ROI feature extraction subnet	ResNet18 with ROI feature extraction subnet	validation set	55.97%

Otberdout et al. [65]	Chehra face tracker [66], resize to 224x224	ExpNet with global covariance descriptor	SVM (Gaussian kernel)	validation set	49.18%
Gogić et al. [53]	-	LBF using ensemble of decision trees	Neural network (1 hidden layer)	validation set	49.31%

Table 7: Results of Reviewed Papers on RAF-DB

Literature	**Pre-processing**	**Feature Extraction**	**Classifier**	**Evaluation Protocol**	**Accuracy**
Acharya et al. [33]	data augmentation	DLP-CNN with covariance pooling	DLP-CNN with covariance pooling	test set	87.00%
Luo et al. [57]	random cropping to 90x90, min-max normalization, data augmentation	VGG-like Deep CNN with Local Subclass Constraint	SVM (RBF kernel)	test set	76.02%
Ji et al. [49]	-	ICID fusion network with DarkNet-19 CNN	DarkNet-19 CNN	test set	75.40%
Tong et al. [36]	data augmentation	DenseNet121 with covariance pooling	DenseNet121 with covariance pooling	test set	88.63%
Sadeghi and Raie [63]	face detection, geometric and illumination normalization	novel Local Chi-squared Metric Learning algorithm	SVM	test set	80.74%
Ruan et al. [23]	face detector, random cropping to 90x90 pixels, data augmentation	pre-trained novel Disturbance Feature Extraction Model (DFEM)	pre-trained novel Disturbance Feature Extraction Model (DFEM)	test set	87.71%

Ly et al. [34]	MTCNN face detector	Ensemble CNNs (Inception-ResNet V1 and PointCNN)	SVM	test set	56.20%
Ly et al. [35]	MTCNN face detector	Ensemble CNNs (Inception-ResNet V1 and PointCNN)	SVM	test set	56.20%
Sadeghi and Raie [40]	face and facial landmark detection (One millisecond), normalization to 150x110, grayscale conversion	Gabor filter	SVM (polynomial kernel)	10-fold cross-validation	76.23%

Table 5 summarizes the results of the reviewed works in this systematic review. It is worth noting that some works tested their algorithms using different evaluation protocols instead of having a commonly agreed convention. As such, the performance comparison is only suitable for those works that use the same validation method.

Regarding a decisional aspect in striving to improve classification accuracy, this review work reveals that those efforts can be generally classified into several strategies: apply more pre-processing techniques such as face detection, face alignment, and data augmentation; putting more effort into the feature extraction stage by using ensemble model to extract deep features, or novel expression-specific features, or techniques that improve feature extraction including covariance pooling and attention block; combine deep features with ML classifier such as SVM; and ensemble classifier that consist of multiple CNNs. In fact, researchers often exploit multiple strategies in their research. The following part discusses the merits and demerits of each strategies as well as relevant findings from reviewed papers.

A face detector is used for locating ROI from face images to facilitate further feature extraction. The Viola-Jones face detector is prevalent in reviewed works because of its fast and accurate detection. However, it degrades significantly in detecting non-frontal faces and complex facial images which commonly exists in unconstrained datasets because it works best for fully frontal upright faces that are often found in the posed dataset. Therefore some research prefers MTCNN as it comes with both face detection and alignment while maintaining real-time performance. The possible reason for applying face detection for SFEW and RAF-DB could be the face region is sometimes not located at the center and occupied small spaces of such big-size images at 100x100 pixels and above. Regarding to FER-2013, not much research has applied a face detector probably because the face is approximately centered and occupies the same area in a much smaller image size at 48x48 pixels. Therefore, it makes sense that a face detector could effectively improve the classification accuracy for SFEW and RAF-DB datasets.

Data augmentation is the most preferred in reviewed works especially research on SFEW and RAF-DB. This is because data augmentation can effectively address the over-fitting problem due to insufficient training data by artificially generating more training data. Compared to the FER-2013 dataset, SFEW and RAF-DB have less amount of training data at 847 images and 12,271 images respectively, while FER-2013 has a massive amount of 28,709 training images. This explains the popularity of data augmentation in the works that involved SFEW and RAF-DB but were less favored for research using the FER-2013 dataset. Despite that, data augmentation might change the original feature of the data hence decreasing the classification accuracy if excessive augmentation is applied [67]. Besides data augmentation, other pre-processing techniques such as Isotropic Smoothing and Histogram Equalization are useful in dealing with varying illumination in unconstrained datasets, although they may not show significant improvement in classification accuracy. Therefore, more work is required to understand their impact.

Numerous reviewed works attempt to refine the feature extraction. Acharya et al. [33] shows that covariance pooling computed from features is better in representing regional features than first-order information from ordinary max or average pooling. The author concludes that this kind of second-order network is more effective in capturing facial landmark distortions. Similarly, Tong et al. [36] shows second-order pooling methods yield 2-4% higher accuracy on RAF-DB and 5-10% on SFEW compared to first-order pooling. The proposed method that introduces second-order pooling into DenseNet121 CNN shows significantly better accuracy than classical CNN and faster convergence speed. Besides, the proposed novel Deep Disturbance-disentangled Learning Model in Ruan et al. [23] combines Disturbance Feature Extraction Model (DFEM) and Disturbance-Disentangled Model (DDM) which is a CNN with attention block scored highest and second-best accuracy for SFEW and RAF-DB. Furthermore, Sadeghi and Raie [63]

presents a novel Local Chi-squared Metric Learning (LCML) method for feature extraction, and Wang et al. [64] added attention subnet in ResNet18 CNN for better-extracting ROI features.

Next, certain works fed deep features to ML classifiers instead of using the normal DL-based approach that handles feature extraction to classification. In Ly et al. [34] and Ly et al. [35] the author achieved that by replacing the softmax layer with SVM as the classifier, while Luo et al. [57] use output from the softmax layer of VGG CNN as input to train SVM by stacking SVM at the last layer of CNN. Both techniques show a slight increase in accuracy compared to the regular softmax CNN model. This strategy prevents the laborious manual feature extraction process in the conventional approach and improves interpretability in the DL-based approach. However, this strategy is more complex and the performance is determined by interrelated factors from both conventional and DL-based approaches. Besides these strategies, ensemble size is equally important in designing an ensemble model as bigger ensemble sizes compromise both training and classification speed and are prone to over-fitting, while smaller ensemble sizes tend to fall in local optimum. For FER-2013, the author concluded a few key points from their extensive investigation: shallow CNN is more effective as a base classifier due to higher sensitivity, simple average voting is a good choice, performance does not increase linearly with the number of base classifiers, shuffling of training data and initial weight does not provide enough variability.

Ensemble methods with multiple CNNs as base classifiers have offered great performance improvement, particularly for the dataset with highly diverse and massive training data such as FER-2013. Based on the reviewed works, the ensemble classifier yields the best results for FER-2013 [32]. The author revealed the key to the formation of an ensemble is selecting appropriate base classifiers that satisfy both accuracy and diversity. The author further explained four strategies that could be considered to increase the diversity in the base classifier: varying the architecture, varying the algorithms, varying the initial weight, and varying the training data.

From the perspective of evaluation protocol, most reviewed works followed the default training and validating setting of the dataset, although some validate their result using k-fold cross-validation. The possible reason to follow the default setting of the dataset could be a convenient result comparison with prior works which report results only on validation or test set. The choice of evaluation protocol is crucial in determining the actual performance of the experimental output. As a matter of fact, different protocols could lead to different results even for the same algorithm and dataset. The reason to employ k-fold cross-validation as an evaluation protocol is to ensure the robustness and generality of the algorithm by minimizing bias and influence of the uneven distribution of samples in each class, which is also common in unconstrained datasets. Therefore, it is suggested future work should include both default setting and k-fold cross-validation as evaluation protocol although it may introduce extra work, a standard protocol is needed to provide a universal quantitative comparison between the performance of each algorithm.

4.5 Summary of Key Findings

RQ1: FER in-the-wild processes

- Pre-processing: Face detection (Viola-Jones, MTCNN), face alignment, normalization, histogram equalization, data augmentation
- Feature extraction: Handcrafted (HOG, LBP, LPQ, Gabor) and deep learning features (CNN)

- Classification: Conventional ML (SVM, KNN) and deep learning (CNN variants)

RQ2: Most widely used methods

- Face detection: Viola-Jones (8 papers), MTCNN (6 papers)
- Pre-processing: Data augmentation (one-third of papers)
- Features: Deep learning features (half of papers)
- Classification: CNN variants (most popular), SVM (second most popular)

RQ3: Strengths and drawbacks of methods used

- Conventional ML: Better interpretability, works with small datasets, requires manual feature engineering
- Deep learning: End-to-end training, automatic feature extraction, requires large datasets, lacks interpretability
- Ensemble methods: Better accuracy and robustness, more complex and computationally expensive

RQ4: Most widely used datasets

- FER-2013, SFEW, and RAF-DB
- Each addresses different aspects of unconstrained facial expressions

RQ5: Main challenges

- Varying illumination, poses, and occlusions in unconstrained datasets
- Uneven distribution of samples across expression classes
- Ambiguous facial expressions
- Insufficient training data.

5.TRENDS AND CHALLENGES

5.1 Key Trends

- **Shift toward deep learning approaches:** Half of the reviewed papers employed DL-based methods, indicating a somewhat growing preference compared to conventional ML approaches.
- **Second-order feature pooling:** Covariance pooling in CNNs showing 2-4% accuracy improvement on RAF-DB and 5-10% on SFEW over first-order pooling.
- **Hybrid model integration:** $\sim$ 15% of studies combine deep features with ML classifiers like SVM for increased interpretability.
- **Ensemble methods:** Especially effective for diverse datasets like FER-2013, showing highest accuracy (72.25%) and thus gaining popularity.
- **Dataset preferences:** FER-2013, SFEW, and RAF-DB emerged as most widely used FER datasets.
- **Strategic data augmentation:** Particularly useful for smaller datasets like SFEW and

RAF-DB to solve over-fitting issues.

- **Attention mechanisms:** Growing integration of attention blocks in CNN architectures to improve region-of-interest feature extraction.
- **Face detection importance:** MTCNN appear to be more popular compared to Viola-Jones for combined face detection and alignment capabilities.

5.2 Key Challenges

- **Illumination and pose variations:** Major obstacles in unconstrained datasets that can lead to feature extraction difficulties.
- **Non-frontal face detection degradation:** Some face detectors still struggle with complex facial images in uncontrolled environments.
- **Uneven expression class distribution:** Common issue of imbalance in spontaneous expression datasets.
- **Inconsistent evaluation protocols:** Lack of standardization in the evaluation methods which hinder fair cross-study comparisons.
- **Limited interpretability:** Deep learning approaches function as ”black boxes” with less explainability.
- **Ensemble model complexity:** Difficulty in determining optimal ensemble architecture and strategies.
- **Data augmentation balance:** Risk of excessive augmentation thereby causing alteration of original features and subsequently decreasing accuracy.
- **Occlusions and background noise:** More common in the unconstrained datasets, complicating feature extraction processes.
- **Real-world deployment hurdles:** Gap between laboratory performance and real-world application effectiveness.

6. CONCLUSIONS

This review presents a holistic view of prior research on image-based FER on spontaneous expression datasets. A total of 32 papers that were published were analysed and assessed in conducting this systematic literature review. Five research questions are answered by studying and understanding the approaches, techniques, dataset, evaluation protocol, and key issues to provide insights for future research.

DL-based methods seem to be the better choice to produce promising results at this moment, yet conventional methods are getting less popular. Similar to other research fields, DL emerged in the last decades and immediately become dominant, yet this kind of black-block algorithm lacks interpretability and does not work well with small data. Hence, there is a need to put more effort into combining both methods organically to keep only their benefits such as the convenience of deep feature extraction and interpretability of ML classifier. For pre-processing, it is encouraged to apply face detection and data augmentation. Face detection is extremely useful for datasets consisting of large image sizes and small face regions. Data augmentation can effectively increase the diversity of training data and eliminate the over-fitting problem. Many reviewed works have shown the importance of face detection and data augmentation for

better accuracy. Next, CNN architecture should embrace second-order pooling for a better representation of regional features and facial landmark distortion. Several works have validated the significant difference between classical CNN with first-order pooling and CNN with second-order pooling. Ensemble strategies are extensively studied in the reviewed work but it requires more investigation to figure out a clear methodology and understanding of architecture that is suitable in this field. Many issues are observed in spontaneous facial expressions datasets such as non-face images, illumination and pose variation, occlusions, ambiguous face expression, and uneven distribution of samples in each class. Future work is needed to overcome this bottleneck. On the bright side, these adversities could help researchers to build a FER algorithm that faces real-world scenarios.

The findings of this review study have revealed that while DL-based approaches, in particularly ensemble methods with covariance pooling. These approaches deliver superior accuracy for facial expression recognition in unconstrained environments. Meanwhile, conventional ML methods remain valuable for resource-constrained applications that require a certain level of interpretability. Subsequently, pre-processing techniques should be selectively applied based on dataset characteristics. Face detection is beneficial datasets with non-centered faces, data augmentation is crucial for limited training samples, and illumination normalization techniques are critical for real-world deployment that have significantly variable lighting conditions. In short, the information provided in this study can serve as a basis for beginning research in image-based FER for spontaneous facial expressions in the wild. As a result, future studies on the current topic are therefore required in the identified directions to establish a comprehensive understanding.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[2] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: A survey," *Symmetry*, vol. 11, no. 10, p. 1189, 2019.

[3] I. M. Revina and W. S. Emmanuel, "A survey on human face expression recognition techniques," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 6, pp. 619–628, 2021.

[4] J. Kumari, R. Rajesh, and K. Pooja, "Facial expression recognition: A survey," *Procedia computer science*, vol. 58, pp. 486–491, 2015.

[5] Y.-L. Tian, T. Kanade, and J. F. Cohn, *Facial Expression Analysis*. New York, NY: Springer New York, 2005, pp. 247–275. [Online]. Available: https://doi.org/10.1007/0-387-27257-7_12

[6] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, Jan. 2003. [Online]. Available: https://doi.org/10.1016/s0031-3203(02)00052-3

[7] N. Wang, X. Gao, D. Tao, H. Yang, and X. Li, "Facial feature point detection: A comprehensive survey," *Neurocomputing*, vol. 275, pp. 50–65, Jan. 2018. [Online]. Available: https://doi.org/10.1016/j.neucom.2017.05.013

[8] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical*

and Engineering Sciences, vol. 374, no. 2065, p. 20150202, Apr. 2016. [Online]. Available: https://doi.org/10.1098/rsta.2015.0202

[9] Y. Liu, Y. Cao, Y. Li, M. Liu, R. Song, Y. Wang, Z. Xu, and X. Ma, "Facial expression recognition with PCA and LBP features extracting from active facial patches," in *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*. IEEE, Jun. 2016. [Online]. Available: https://doi.org/10.1109/rcar.2016.7784056

[10] A. J. Calder, A. Burton, P. Miller, A. W. Young, and S. Akamatsu, "A principal component analysis of facial expressions," *Vision Research*, vol. 41, no. 9, pp. 1179–1208, Apr. 2001. [Online]. Available: https://doi.org/10.1016/s0042-6989(01)00002-5

[11] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, "Facial expression recognition using a hybrid CNN–SIFT aggregator," in *Lecture Notes in Computer Science*. Springer International Publishing, 2017, pp. 139–149. [Online]. Available: https://doi.org/10.1007/978-3-319-69456-6_12

[12] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. [Online]. Available: https://doi.org/10.1109/5.726791

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2015. [Online]. Available: https://doi.org/10.1109/cvpr.2015.7298594

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. Red Hook, NY, USA: Curran Associates Inc., 2012, p. 1097–1105.

[15] Y. Tang, "Deep learning using linear support vector machines," 2013.

[16] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing*. Springer Berlin Heidelberg, 2013, pp. 117–124. [Online]. Available: https://doi.org/10.1007/978-3-642-42051-1_16

[17] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1. Ieee, 2001, pp. I–I.

[18] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016. [Online]. Available: https://doi.org/10.1109/lsp.2016.2603342

[19] H. Jiang and E. Learned-Miller, "Face detection with the faster r-cnn," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 650–657.

[20] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[21] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3444–3451.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] D. Ruan, Y. Yan, S. Chen, J.-H. Xue, and H. Wang, "Deep disturbance-disentangled learning for facial expression recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, Oct. 2020. [Online]. Available: https://doi.org/10.1145/3394171.3413907

[24] S. Yu, J. Wu, S. Wu, and D. Xu, "Lib face detection," 2016. [Online]. Available: https://github.com/ShiqiYu/libfacedetection

[25] H. Hussein, F. Angelini, M. Naqvi, and J. A. Chambers, "Deep-learning based facial expression recognition system evaluated on three spontaneous databases," in *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC)*. IEEE, Nov. 2018. [Online]. Available: https://doi.org/10.1109/isivc.2018.8709224

[26] S. Arshid, A. Hussain, A. Munir, A. Nawaz, and S. Aziz, "Multi-stage binary patterns for facial expression recognition in real world," *Cluster Computing*, vol. 21, no. 1, pp. 323–331, Mar. 2017. [Online]. Available: https://doi.org/10.1007/s10586-017-0832-5

[27] Z. Fei, E. Yang, D. D.-U. Li, S. Butler, W. Ijomah, X. Li, and H. Zhou, "Deep convolution network based emotion analysis towards mental health care," *Neurocomputing*, vol. 388, pp. 212–227, May 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2020.01.034

[28] S. Xie, H. Hu, and Y. Wu, "Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition," *Pattern Recognition*, vol. 92, pp. 177–191, Aug. 2019. [Online]. Available: https://doi.org/10.1016/j.patcog.2019.03.019

[29] M. S. Zia, M. Hussain, and M. A. Jaffar, "A novel spontaneous facial expression recognition using dynamically weighted majority voting based ensemble classifier," *Multimedia Tools and Applications*, vol. 77, no. 19, pp. 25 537–25 567, Mar. 2018. [Online]. Available: https://doi.org/10.1007/s11042-018-5806-y

[30] S. A. Khan, A. Hussain, and M. Usman, "Reliable facial expression recognition for multi-scale images using weber local binary image based cosine transform features," *Multimedia Tools and Applications*, vol. 77, no. 1, pp. 1133–1165, Jan. 2017. [Online]. Available: https://doi.org/10.1007/s11042-016-4324-z

[31] C. Wang, K. Lu, J. Xue, and Y. Yan, "Dense attention network for facial expression recognition in the wild," in *Proceedings of the ACM Multimedia Asia*. ACM, Dec. 2019. [Online]. Available: https://doi.org/10.1145/3338533.3366568

[32] A. Renda, M. Barsacchi, A. Bechini, and F. Marcelloni, "Comparing ensemble strategies for deep learning: An application to facial expression recognition," *Expert Systems with Applications*, vol. 136, pp. 1–11, Dec. 2019. [Online]. Available: https://doi.org/10.1016/j.eswa.2019.06.025

[33] D. Acharya, Z. Huang, D. P. Paudel, and L. V. Gool, "Covariance pooling for facial expression recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Jun. 2018. [Online]. Available: https://doi.org/10.1109/cvprw.2018.00077

[34] S. T. Ly, N.-T. Do, G.-S. Lee, S.-H. Kim, and H.-J. Yang, "A 3d face modeling approach for in-the-wild facial expression recognition on image datasets," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, Sep. 2019. [Online]. Available: https://doi.org/10.1109/icip.2019.8803434

[35] T. S. Ly, N.-T. Do, S.-H. Kim, H.-J. Yang, and G.-S. Lee, "A novel 2d and 3d multimodal approach for in-the-wild facial expression recognition," *Image and Vision Computing*, vol. 92, p. 103817, Dec. 2019. [Online]. Available: https://doi.org/10.1016/j.imavis.2019.10.003

[36] X. Tong, S. Sun, and M. Fu, "Data augmentation and second-order pooling for facial expression recognition," *IEEE Access*, vol. 7, pp. 86 821–86 828, 2019. [Online]. Available: https://doi.org/10.1109/access.2019.2923530

[37] J. Shao and Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild," *Neurocomputing*, vol. 355, pp. 82–92, Aug. 2019. [Online]. Available: https://doi.org/10.1016/j.neucom.2019.05.005

[38] S. Wang, Y. Yuan, X. Zheng, and X. Lu, "Local and correlation attention learning for subtle facial expression recognition," *Neurocomputing*, Sep. 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2020.07.120

[39] G. V. Reddy, C. D. Savarni, and S. Mukherjee, "Facial expression recognition in the wild, by fusion of deep learnt and hand-crafted features," *Cognitive Systems Research*, vol. 62, pp. 23–34, Aug. 2020. [Online]. Available: https://doi.org/10.1016/j.cogsys.2020.03.002

[40] H. Sadeghi and A.-A. Raie, "Human vision inspired feature extraction for facial expression recognition," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 30 335–30 353, Jun. 2019. [Online]. Available: https://doi.org/10.1007/s11042-019-07863-z

[41] F. Zhang, T. Zhang, Q. Mao, L. Duan, and C. Xu, "Facial expression recognition in the wild," in *Proceedings of the 26th ACM international conference on Multimedia*. ACM, Oct. 2018. [Online]. Available: https://doi.org/10.1145/3240508.3240574

[42] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2852–2861.

[43] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 2106–2112.

[44] A. Dapogny, K. Bailly, and S. Dubuisson, "Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 255–271, Apr. 2017. [Online]. Available: https://doi.org/10.1007/s11263-017-1010-1

[45] M. H. Siddiqi, M. Ali, M. E. A. Eldib, A. Khan, O. Banos, A. M. Khan, S. Lee, and H. Choo, "Evaluating real-life performance of the state-of-the-art in facial expression recognition using a novel YouTube-based datasets," *Multimedia Tools and Applications*, vol. 77, no. 1, pp. 917–937, Jan. 2017. [Online]. Available: https://doi.org/10.1007/s11042-016-4321-2

[46] W. Sun, H. Zhao, and Z. Jin, "A visual attention based ROI detection method for facial expression recognition," *Neurocomputing*, vol. 296, pp. 12–22, Jun. 2018. [Online]. Available: https://doi.org/10.1016/j.neucom.2018.03.034

[47] C. Turan and K.-M. Lam, "Histogram-based local descriptors for facial expression recognition (FER): A comprehensive study," *Journal of Visual Communication and Image Representation*, vol. 55, pp. 331–341, Aug. 2018. [Online]. Available: https://doi.org/10.1016/j.jvcir.2018.05.024

[48] C. Wang, S. Wang, and G. Liang, "Identity- and pose-robust facial expression recognition through adversarial feature learning," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, Oct. 2019. [Online]. Available: https://doi.org/10.1145/3343031.3350872

[49] Y. Ji, Y. Hu, Y. Yang, F. Shen, and H. T. Shen, "Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network," *Neurocomputing*, vol. 333, pp. 231–239, Mar. 2019. [Online]. Available: https://doi.org/10.1016/j.neucom.2018.12.037

[50] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," *Procedia Computer Science*, vol. 175, pp. 689–694, 2020. [Online]. Available: https://doi.org/10.1016/j.procs.2020.07.101

[51] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," in *International workshop on analysis and modeling of faces and gestures*. Springer, 2007, pp. 168–182.

[52] A. T. Lopes, E. de Aguiar, A. F. D. Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, Jan. 2017. [Online]. Available: https://doi.org/10.1016/j.patcog.2016.07.026

[53] I. Gogić, M. Manhart, I. S. Pandžić, and J. Ahlberg, "Fast facial expression recognition using local binary features and shallow neural networks," *The Visual Computer*, vol. 36, no. 1, pp. 97–112, Aug. 2018. [Online]. Available: https://doi.org/10.1007/s00371-018-1585-8

[54] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[55] A. Munir, A. Hussain, S. A. Khan, M. Nadeem, and S. Arshid, "Illumination invariant facial expression recognition using selected merged binary patterns for real world images," *Optik*, vol. 158, pp. 1016–1025, Apr. 2018. [Online]. Available: https://doi.org/10.1016/j.ijleo.2018.01.003

[56] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. [Online]. Available: https://doi.org/10.1038/nature14539

[57] Z. Luo, J. Hu, and W. Deng, "Local subclass constraint for facial expression recognition in the wild," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, Aug. 2018. [Online]. Available: https://doi.org/10.1109/icpr.2018.8545847

[58] P. Saha, D. Bhattacharjee, B. K. De, and M. Nasipuri, "A survey on image acquisition protocols for non-posed facial expression recognition systems," *Multimedia Tools and Applications*, vol. 78, no. 16, pp. 23 329–23 368, May 2019. [Online]. Available: https://doi.org/10.1007/s11042-019-7596-2

[59] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 5 pp.–.

[60] F. WALLHOFF, "The facial expressions and emotions database homepage (feedtum)," *www.mmk.ei.tum.de/∼waf/fgnet/feedtum.html*, 2005. [Online]. Available: https://ci.nii.ac.jp/naid/10019625286/en/

[61] C. E. Erdem, C. Turan, and Z. Aydin, "BAUM-2: a multilingual audio-visual affective face database," *Multimedia Tools and Applications*, vol. 74, no. 18, pp. 7429–7459, May 2014. [Online]. Available: https://doi.org/10.1007/s11042-014-1986-2

[62] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, Jan. 2019. [Online]. Available: https://doi.org/10.1109/taffc.2017.2740923

[63] H. Sadeghi and A.-A. Raie, "Histogram distance metric learning for facial expression recognition," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 152–165, Jul. 2019. [Online]. Available: https://doi.org/10.1016/j.jvcir.2019.05.004

[64] C. Wang, K. Lu, J. Xue, and Y. Yan, "R-FENet: A region-based facial expression recognition method inspired by semantic information of action units," in *Proceedings of the 1st International Workshop on Human-centric Multimedia Analysis*. ACM, Oct. 2020. [Online]. Available: https://doi.org/10.1145/3422852.3423482

[65] N. Otberdout, A. Kacem, M. Daoudi, L. Ballihi, and S. Berretti, "Automatic analysis of facial expressions based on deep covariance trajectories," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 3892–3905, Oct. 2020. [Online]. Available: https://doi.org/10.1109/tnnls.2019.2947244

[66] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1859–1866.

[67] D. Canedo and A. J. R. Neves, "Facial expression recognition using computer vision: A systematic review," *Applied Sciences*, vol. 9, no. 21, p. 4678, Nov. 2019. [Online]. Available: https://doi.org/10.3390/app9214678