An In-Depth Review of Deep Facial Expression Recognition: Obstacles, Utilizations, and Prospective Directions

1Raed Ibrahim Khaleel Almsari, 2Abbas Hussein Miry, 3Tariq M. Salman

1MSC Student, Electrical Engineering Department, Al-Mustansiriya University, Baghdad, Iraq
2,3Assistant Professor Dr., Electrical Engineering Department, Al-Mustansiriya University, Baghdad, Iraq

Abstract - Facial expression recognition (FER) is emerging as an emerging and multifaceted field of study. The use of FER in areas such as healthcare, security, and safe driving has not only enhanced the credibility of these technologies, but also their integration into human computer interaction to achieve intelligent outcomes. Computational FER seeks to replicate the skill of humans in decoding facial expressions, providing important cues that complement spoken language and aid listeners in understanding. Likewise, FER's deep learning (DL) and Artificial Intelligence (AI) methodologies are meticulously designed, incorporating advanced modules for efficiency and real time processing. In light of this background, many investigations have looked at different aspects of FER. Although current surveys focus primarily on traditional technologies and generic methodologies for on premises servers, they overlook the large field of deep learning inspired by edge vision and AI assisted FER technologies. To fill this gap, the current study conducts a comprehensive and thorough analysis of the prevailing FER literature. It carefully surveys the operational framework of FER technologies, highlighting their basic and intermediate phases, as well as the underlying pattern structures. Furthermore, the study addresses the limitations inherent in current FER surveys. The exploration extends to the FER datasets, subjecting them to thorough examination, thus revealing the attendant challenges and pitfalls. In addition, it provides a comprehensive discussion of the various metrics used to measure the effectiveness of FER methods.

Keywords: Facial Emotion Recognition; Dataset; Deep Learning; Computer Vision.

I. INTRODUCTION

FER serves a myriad of purposes, spanning from human computer interaction intelligence and eLearning systems to marketing, stress analysis, and diverse interactive services. Such services, in order to remain pertinent, necessitate real time status updates. Thus, real time FER processing assumes a momentous role within the domain of computer vision. Amidst this landscape, the utilization of convolutional neural network (CNN) models for FER has garnered considerable attention, primarily fueled by the leverage of extensive datasets for DL. While various survey papers have meticulously documented the evolution of FER techniques [3]. Researchers in [4] have also probed into real time FER methodologies. However, a noticeable void exists in the form of survey papers covering FER that harnesses few shot learning (FSL) technology, which stands as an alternative learning strategy. This technique exhibits the potential to address prevailing challenges in deep learning.

This study delves into recent FER systems that capitalize on FSL, a fact that bears the potential to mitigate existing challenges in the FER domain. Bringing matters to a close, the final section encapsulates the conclusions drawn from this comprehensive review. The uniqueness of this endeavor stems from the ensuing aspects:

- This denotes the debut survey paper focused on FER carried out through FSL procedures.
- We offer a thorough assessment of contemporary profound learning models applied to FER, focusing on the difficulties they wrestle with, deliberately arranging them in view of the idea of information issues and philosophies.
- We dive into the domain of FER through FSL by comparing two particular situations: the speculation to novel information and area transformation.
- We feature the potential for FER involving FSL to present time effectiveness and lessen unpredictability in various constant picture handling attempts.
II. FER DATASETS

Numerous datasets catering to facial expression recognition (FER) are at our disposal within the FER domain. Nevertheless, our focus narrows down to a select subset of datasets, particularly those tailored for FER through FSL methodologies. In the realm of deep learning driven FER, the acquisition of an optimal volume of data stands as a pivotal challenge. This stems from the considerable variations that facial images exhibit due to factors like age, gender, and ethnicity. Datasets can be neatly bifurcated into 'in the lab' and 'in the wild' categories. An 'in the lab' dataset typically furnishes pristine, high quality data alongside meticulously structured and annotated emotional cues, often mapped through the Facial Action Code, a system tailored for controlled environments. Conversely, 'in the wild' datasets encapsulate spontaneous facial expressions captured across diverse environmental contexts.

FER datasets can be further categorized into image based and video based datasets. The latter, characterized by sequences of emotions, enables a more comprehensive information accrual compared to single emotion images.

The onset of initiatives like the Emotion Recognition in the Wild Challenge (EmotiW) has facilitated the access to a plethora of unconstrained facial images sourced from the digital realm. However, it's noteworthy that FER models formulated within controlled environments often demonstrate performance degradation when applied to the real world, where unscripted and sequential images prevail. Consequently, a thrust has been placed on the exploration of deep learning-based models to harness real world datasets.

Within this segment, we delve into diverse datasets pertinent to FSL centric FER tasks. The FER 2013 dataset, FER+ [9], AffectNet [5], Moreover, the Reality Full of Emotion Faces Dataset (RAFDB) [13] comprises images depicting six primary emotions (anger, disgust, fear, joy, sadness, and surprise), accompanied by a neutral sentiment. In contrast, the extensive Cohn-Kanade (CK+) dataset comprises seven fundamental emotions (anger, contempt, disgust, fear, joy, sadness, and surprise). Additionally, considering the challenge of domain shift assumption frequently faced in Few-Shot Learning (FSL), several experiments involve training on datasets containing basic emotions and subsequently evaluating on datasets containing complex emotions. As a result, we also delve into the RAFDB and the Compound Looks of Emotion (CFEE) dataset [11], both encompassing a spectrum of compound emotions such as 'sadly surprised' and 'fearfully angry'. Table 1 summarizing the characteristics of each dataset.

**Table 1: Emotion datasets summary**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Contents</th>
<th>Emotion Categories</th>
<th>Data Collection</th>
<th>Annotations</th>
<th>Size</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CK+ [10]</td>
<td>593 sequences 123 subjects</td>
<td>7 basic emotions + contempt</td>
<td>Lab controlled</td>
<td>Neutral to peak expression</td>
<td>Medium</td>
<td>Widely used for FER evaluation</td>
</tr>
<tr>
<td>FER2013 [8]</td>
<td>35,887 images 3 groups</td>
<td>6 basic expressions + neutral</td>
<td>Google API &amp; OpenCV</td>
<td>Human labelers</td>
<td>Large</td>
<td>Training, validation, testing</td>
</tr>
<tr>
<td>FER+ [9]</td>
<td>Improved FER dataset</td>
<td>Standard emotions + probabilities</td>
<td>Crowdsourced taggers</td>
<td>10 labelers per image</td>
<td>Medium</td>
<td>Multi-label algorithms</td>
</tr>
<tr>
<td>EmotioNet [12]</td>
<td>1 million facial images</td>
<td>Various expressions</td>
<td>AU detection model</td>
<td>Automatic &amp; manual annotation</td>
<td>Large</td>
<td>AU detection model evaluation</td>
</tr>
<tr>
<td>AffectNet [2]</td>
<td>1,000,000+ facial images</td>
<td>8 basic expressions</td>
<td>Internet search</td>
<td>Manual annotation</td>
<td>Large</td>
<td>Categorical &amp; dimensional models</td>
</tr>
<tr>
<td>RAFDB [13]</td>
<td>29,672 diverse images</td>
<td>Common expressions + compound</td>
<td>Internet download</td>
<td>Not specified</td>
<td>Medium</td>
<td>Real world expressions</td>
</tr>
<tr>
<td>AFEW [14]</td>
<td>957 videos</td>
<td>6 basic expressions + neutral</td>
<td>Collected from movies</td>
<td>Manual annotation</td>
<td>Small</td>
<td>Challenging environment analysis</td>
</tr>
</tbody>
</table>
III. FACE FEELING ACKNOWLEDGMENT

TECHNIQUES AND ADVANCES

The process of identifying facial emotions involves a series of distinct stages, starting from face detection and culminating in emotion classification. Initially, face detection involves isolating the face from its surroundings through the identification of key facial features such as the eyes, nose, and mouth. Once this step is completed, the system progresses to the task of categorizing the emotions [15].

3.1 Face Detection

Face detection is a foundational concept in computer vision that involves identifying and locating human faces within images or video frames [16]. This process relies on algorithms that analyze the visual patterns and unique attributes associated with human faces, such as the arrangement of eyes, nose, mouth, and other facial features. Effective face detection is essential for various applications, including facial recognition, emotion analysis, and even in digital cameras for focusing and exposure adjustment. Modern face detection methods often utilize machine learning techniques, particularly convolutional neural networks (CNNs), to achieve high accuracy and robustness in detecting faces across different poses, lighting conditions, and backgrounds. [17][18].

3.2 Feature Extraction

Face image feature extraction is a crucial process in computer vision that involves distilling meaningful and distinctive information from facial images. It aims to capture key characteristics such as facial expressions, shapes, textures, and structures. By converting complex facial data into a more compact and representative form, feature extraction facilitates efficient analysis and recognition tasks. Various techniques, including deep learning approaches like convolutional neural networks (CNNs) and traditional methods like Principal Component Analysis (PCA), are employed to extract these features. The extracted features play a vital role in applications ranging from facial emotion recognition and identity verification to surveillance and human-computer interaction. [15].

3.3 Data Augmentation

Data augmentation aims to improve emotion classification outcomes by enhancing the dataset. Common practices include altering contrast and brightness, as well as resizing images to a uniform format. Additional techniques involve filtering and edge detection, often using methods like the Sobel edge detection [17].

3.4 Emotion Classification

Emotion classification is a significant aspect of both psychology and artificial intelligence, aiming to decipher and categorize human emotional states based on various cues such as facial expressions, voice tone, and physiological signals. In the realm of AI, emotion classification involves training models to recognize and differentiate between different emotional states, such as happiness, sadness, anger, fear, and surprise, among others [19]. This is achieved through supervised learning techniques, where machine learning algorithms learn from labeled datasets containing examples of different emotions. These models are then capable of analyzing input data, such as images or audio recordings, and assigning them to appropriate emotion categories. Emotion classification finds applications in diverse fields including human-computer interaction, sentiment analysis, mental health monitoring, and the development of empathetic AI systems [20].

3.5 Real Time Emotion Recognition

Real-time emotion recognition is a dynamic technology that involves the instantaneous detection and interpretation of human emotional states as they occur. Utilizing various data sources such as facial expressions, speech patterns, and physiological signals, real-time emotion recognition systems employ sophisticated algorithms to rapidly analyze and categorize emotions like happiness, sadness, anger, and surprise [21]. This technology holds immense potential in applications like video conferencing, virtual reality, and human-computer interaction, where understanding and responding to users' emotions in real time can enhance user experience and engagement. Real-time emotion recognition is challenging due to the need for speed and accuracy, often requiring advanced machine learning models and efficient processing pipelines to ensure seamless and meaningful emotion interpretation in dynamic contexts [22].

IV. STATE OF THE ART TECHNOLOGIES, MEASURES, AND MODELS

Several cutting-edge technologies, measures, and models contribute to advanced emotion recognition systems. Table 2 describing each of the mentioned technologies, methods, and models.
Table 2: Advanced technologies, methods, and models for facial emotion recognition

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Key Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAERNet</td>
<td>A model for context aware emotion recognition that considers both facial and contextual cues [23].</td>
<td>Focuses on face and attentive context regions. Enhances emotion recognition with contextual information.</td>
</tr>
<tr>
<td>SASEFE</td>
<td>Dataset featuring congruent and incongruent facial expressions, revealing insights into various emotional types [23].</td>
<td>Provides congruent and incongruent expressions. Offers understanding of complex emotional types.</td>
</tr>
<tr>
<td>Microsoft Hololens</td>
<td>Hardware device with a depth camera for 3D face detection, utilizing the Microsoft Azure Face API application [23].</td>
<td>Employs depth camera for accurate 3D detection. Integrates with Azure Face API for enhanced functionality.</td>
</tr>
<tr>
<td>Multimodal Systems</td>
<td>Integrates image, video, and voice inputs for improved emotion determination accuracy [24].</td>
<td>Combines various input modalities for comprehensive analysis. Increases accuracy by leveraging multiple sources of data.</td>
</tr>
<tr>
<td>Stationary Wavelet Entropy</td>
<td>Extracts facial expression features using stationary wavelet entropy [25].</td>
<td>Applies wavelet based method for feature extraction. Captures distinctive features from expressions.</td>
</tr>
<tr>
<td>Island Loss</td>
<td>Technique enhancing discriminative features by reducing intra class variations and increasing interclass differences [11].</td>
<td>Focuses on enhancing feature separability. Improves emotion representation within features.</td>
</tr>
<tr>
<td>Deeply Supervised CNN</td>
<td>Merges models like Feed Forward Neural Network and Naive Bayes for heightened accuracy [15].</td>
<td>Combines different models for improved classification. Enhances accuracy through collaborative learning.</td>
</tr>
<tr>
<td>Facial Expression Sentence</td>
<td>Generates descriptive sentences to interpret recognized facial emotions [26].</td>
<td>Converts emotion recognition into human in temper table language. Provides comprehensive descriptions of emotional states.</td>
</tr>
<tr>
<td>Generating Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compound Emotion Recognition</td>
<td>Identifies complex emotions formed by combining basic emotions [27].</td>
<td>Detects and labels multifaceted emotional states. Expands emotional recognition beyond basic categories.</td>
</tr>
</tbody>
</table>

V. TRANSFER LEARNING IN FER

Face emotion recognition based on transfer learning is a powerful and widely adopted approach in the field of computer vision and affective computing. This methodology leverages pre-trained deep neural networks, often originating from large-scale image classification tasks, to extract meaningful features from facial images and subsequently infer emotional states. Transfer learning capitalizes on the knowledge acquired by these pre-trained models and adapts it to the specific task of emotion recognition [28].

The process typically involves fine-tuning the pre-trained model on a dataset that contains labeled facial expressions, where each image is associated with a particular emotion such as happiness, sadness, anger, fear, disgust, or surprise. During fine-tuning, the weights of the network are adjusted to optimize its performance on the emotion recognition task. This allows the model to learn relevant facial features, such as the arrangement of facial landmarks, the shape of the mouth, or the intensity of eye expressions, that are indicative of different emotional states [29].

One key advantage of transfer learning in face emotion recognition is the ability to achieve high accuracy even when dealing with limited training data, as the model has already learned general features from a vast amount of data during pre-training. This reduces the risk of overfitting and can lead to more robust emotion recognition systems. Additionally, transfer learning enables researchers and developers to avoid the computational cost and time associated with training deep neural networks from scratch, making it a practical and efficient approach for real-world applications such as human-computer interaction, sentiment analysis, and healthcare [30].

However, the choice of the pre-trained model, the fine-tuning strategy, and the selection of an appropriate emotion recognition dataset all play crucial roles in the success of this approach. The field continues to evolve, with ongoing research aimed at improving the performance, robustness, and generalization capabilities of transfer learning-based face emotion recognition systems, making it an exciting area of study with significant potential for various applications in both academia and industry [31].

VI. RELATED WORKS FER

The field of facial expression recognition has witnessed significant advancements through a multitude of studies focusing on various datasets, feature extraction techniques,
and classification methods. Notably, in the work presented by authors [32], they employed the Local Fisher Discriminant Analysis (LFDA) for feature extraction from the JAFFE and MUG datasets, achieving remarkable recognition rates of 94.37% and 95.24%, respectively, using a 1-nearest-neighbor classification approach. In a similar vein, researchers [33] concentrated solely on the JAFFE dataset, utilizing Gabor filters for feature extraction and Bayesian classification techniques, yielding an impressive recognition rate of 96.73%.

Extending the scope to multiple datasets, [34] harnessed Gabor techniques for both JAFFE and Yale datasets, coupling them with a neural network back-propagation algorithm. This hybrid approach led to recognition rates of 96.83% for JAFFE and 92.22% for Yale.

Furthermore, the application of diverse feature extraction methods remained a key avenue of exploration. For instance, [35] delved into the combination of Gabor wavelet transform, Principal Component Analysis (PCA), and Local Binary Pattern (LBP) for feature extraction from JAFFE, demonstrating a recognition rate of 90% using the k-Nearest-Neighbor (k-NN) classification technique. A departure from image-based techniques, [36] focused on the CK+ dataset and introduced kernel PCA (KPCA) for feature extraction, attaining recognition rates of 76.5% and 72.3% for KPCA and PCA, respectively.

Classifiers also played a pivotal role in achieving robust recognition performance. [37] adopted an Eigen face approach with Euclidean distance as the classification metric, resulting in an average recognition rate of 85.38%. In contrast, [38] explored Active Shape Models (ASM) in conjunction with RBF kernel Support Vector Machines (SVM) and Hidden Markov Models (HMM), achieving recognition rates of 70.6% and 65.2%, respectively. Noteworthy is the innovative use of Biorthogonal Wavelet Entropy (BWE) by [39], coupled with Fuzzy Multiclass SVM (FMSVM), attaining an impressive recognition rate of 96.77% ± 0.10%.

Incorporating audio signals into facial expression recognition, [40] combined Gabor filtering for images and Mel-Frequency Cepstral Coefficients (MFCC) for audio, achieving recognition rates of 84.68% for the CK dataset, 80.68% for the Berlin dataset, and 81.58% in real-time scenarios using SVM. In the era of deep learning, [41] employed Convolutional Neural Networks (CNN) to process the JAFFE and CK+ datasets, achieving recognition rates of 76.7442% and 80.303%, respectively.

Additionally, [42] introduced a hybrid feature set consisting of Gabor and Local Binary Pattern (LBP), and examined its effectiveness with various kernel SVMs, achieving impressive recognition rates ranging from 94.45% to 97.42% across different SVM kernels and class distributions. [43] further expanded the study across multiple datasets—CK+, JAFFE, and BU-3DFE—attaining recognition rates of 96.76%, 82.10%, and 82% respectively, through CNN-based approaches.

Even novel techniques like Discrete Wavelet Transform (DWT) found their place in facial expression recognition, as demonstrated by [44], who employed a Single-hidden-layer Neural Network (NN) for classification and achieved a recognition rate of 89.49% ± 0.76%.

### Table 3: Related works summary

<table>
<thead>
<tr>
<th>Paper Reference</th>
<th>Dataset</th>
<th>Feature Extraction</th>
<th>Classification</th>
<th>Recognition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>[32]</td>
<td>JAFFE, MUG</td>
<td>LFDA</td>
<td>1-NN</td>
<td>JAFFE: 94.37%, MUG: 95.24%</td>
</tr>
<tr>
<td>[33]</td>
<td>JAFFE</td>
<td>Gabor filter</td>
<td>Bayesian</td>
<td>96.73%</td>
</tr>
<tr>
<td>[34]</td>
<td>JAFFE, Yale</td>
<td>Gabor techniques</td>
<td>Neural network</td>
<td>JAFFE: 96.83%, Yale: 92.22%</td>
</tr>
<tr>
<td>[35]</td>
<td>JAFFE</td>
<td>Gabor wavelet, PCA, LBP</td>
<td>k-NN</td>
<td>90%</td>
</tr>
<tr>
<td>[36]</td>
<td>CK+</td>
<td>KPCA</td>
<td>KPCA, PCA</td>
<td>KPCA: 76.5%, PCA: 72.3%</td>
</tr>
<tr>
<td>[37]</td>
<td>Private</td>
<td>Eigen face</td>
<td>Euclidean distance</td>
<td>Avg: 85.38%</td>
</tr>
<tr>
<td>[38]</td>
<td>CK+</td>
<td>Active Shape Models</td>
<td>SVM, HMM</td>
<td>SVM: 70.6%, HMM: 65.2%</td>
</tr>
<tr>
<td>[39]</td>
<td>Private</td>
<td>Biorthogonal Wavelet Entropy</td>
<td>Fuzzy Multiclass SVM</td>
<td>96.77% ± 0.10%</td>
</tr>
<tr>
<td>[40]</td>
<td>CK, Berlin</td>
<td>Gabor, MFCC</td>
<td>SVM</td>
<td>CK: 84.68%, Berlin: 80.68%, Real-Time: 81.58%</td>
</tr>
<tr>
<td>[41]</td>
<td>JAFFE, CK+</td>
<td>CNN</td>
<td>CNN</td>
<td>JAFFE: 76.7442%, CK+: 80.303%</td>
</tr>
</tbody>
</table>
The comprehensive survey on deep facial expression recognition provides valuable insights into the rapidly evolving field and highlights various facets of the subject. The study begins by acknowledging the increasing significance of facial expression recognition (FER) in diverse domains, such as healthcare, security, and human-computer interaction. The integration of FER with deep learning and artificial intelligence techniques has propelled its capabilities, enabling real-time processing and intelligent outcomes. The survey underlines the importance of computational FER as a means of replicating human abilities in decoding facial expressions, offering essential cues that complement verbal communication and enhance understanding.

The survey brings to light the limitations of existing FER surveys, which often focus on traditional methods and technologies, disregarding the vast potential of deep learning and AI-assisted FER. To address this gap, the study extensively reviews the current literature on FER, taking into account the operational framework, phases, pattern structures, and challenges associated with FER technologies. This encompasses both the basic and intermediate stages of FER processes, enabling a holistic understanding of the field’s advancements.

Furthermore, the survey dives into the realm of FER datasets, providing a thorough examination of the datasets commonly used in FER tasks. The challenges posed by variations in facial images due to factors like age, gender, and ethnicity are acknowledged. The study categorizes datasets into "inthelab" and "inthewild" categories, highlighting the distinctions between controlled and real-world environments. Moreover, the significance of multimodal datasets that incorporate image, video, and voice inputs is emphasized for improving accuracy.

The survey meticulously explores the technologies, methods, and models employed in FER, offering insights into face detection, feature extraction, data augmentation, emotion classification, and real-time FER. Notably, deep learning frameworks like Convolutional Neural Networks (CNNs) are identified as pivotal for their ability to analyze patterns within data and excel in image recognition tasks. The discussion on advanced technologies further underlines the importance of context-aware emotion recognition, multimodal systems, and novel approaches to handling complex and compound emotions.

A significant contribution of the survey lies in its focus on FER through few-shot learning (FSL), a relatively less-explored area in the existing literature. The study elucidates the potential of FSL to address challenges in deep learning-based FER and presents a comprehensive review of recent FER systems that leverage FSL techniques. This examination sheds light on the versatility of FSL in generalizing to novel data and adapting to different domains, thus improving the efficiency and applicability of FER methods.

**REFERENCES**


Citation of this Article: