**IAES Journal of Electrical and Electronics Engineering**
Vol. 7, No. 2, December 2020, pp. 29~32
ISSN: 2338-7432, DOI: 10.11591/jeee.v7i2.pp29-32 

# Facial expression recognition and deep learning: architectures, datasets, real-world challenges, and future directions

**Anggit Pamungkas[1], Tri Wahono[2], Eneng Nuraeni[2]**
[1]Embedded Systems and Power Electronics Research Group, Yogyakarta, Indonesia
[2]Institute of Advanced Engineering and Science, Yogyakarta, Indonesia

**Article Info**

***Article history:***

Received Mar 20, 2019
Revised Oct 19, 2019
Accepted Jan 11, 2020

***Keywords:***

Affective computing
Computer vision
Convolutional neural networks
Deep learning
Emotion analysis
Facial datasets
Facial expression recognition

**ABSTRACT**

Facial expression recognition (FER) is an important area of computer vision that seeks to characterize visible facial movements from images and video. Deep learning has transformed FER by reducing dependence on handcrafted features and learning hierarchical representations directly from data. However, reliable recognition remains difficult outside controlled laboratory settings because facial appearance changes with illumination, head pose, occlusion, age, identity, and cultural context. This review examines deep-learning approaches to FER, focusing on convolutional neural networks (CNNs), recurrent and temporal models, attention mechanisms, transfer learning, and multimodal systems. It also reviews benchmark datasets, annotation strategies, evaluation protocols, and applications in human–computer interaction, affective computing, and behavioral analysis. The literature shows that larger in-the-wild datasets, including AffectNet and RAF-DB, have improved evaluation realism, while crowdsourced labels and class imbalance continue to introduce uncertainty. Strong benchmark accuracy therefore does not necessarily indicate reliable interpretation of spontaneous expressions or a person's internal emotional state. The review identifies cross-dataset generalization, demographic fairness, explainability, privacy, and ethical deployment as central research priorities. It concludes that robust FER requires technically sound models and careful interpretation of what facial behavior can legitimately reveal.

*This is an open access article under the CC BY-SA license.*

## 1. INTRODUCTION

Facial expressions are important components of non-verbal communication because movements of the eyes, eyebrows, mouth, and other facial regions can convey socially meaningful information. Automatic facial expression recognition (FER) attempts to identify patterns in these movements using computer vision and machine learning. The technology has attracted interest in human–computer interaction, social robotics, behavioral research, driver monitoring, and interactive applications. Nevertheless, translating visible facial movements into reliable computational categories remains a challenging task because expressions vary considerably between individuals and situations [1]–[3].

Earlier FER systems generally relied on handcrafted features, including facial landmarks, geometric measurements, local binary patterns, and texture descriptors. These features were subsequently processed by classifiers such as support vector machines and other conventional machine-learning algorithms. Although these approaches achieved useful results in controlled environments, their performance often deteriorated when images contained illumination changes, head rotation, occlusion, or spontaneous facial movements. Deep learning has changed this research direction by allowing models to learn hierarchical facial representations directly from training data [1], [2], [4].

This review examines the development of deep-learning-based FER up to 2019. It focuses on network architectures, datasets, temporal and multimodal analysis, evaluation practices, and practical limitations. The discussion follows the research problem from uncontrolled facial-image acquisition to representation learning, expression classification, and real-world deployment. A central consideration

throughout the review is that a predicted facial-expression category should not automatically be interpreted as a definitive measurement of a person's internal emotional state.

## 2. RESEARCH PROBLEMS IN FACIAL EXPRESSION RECOGNITION

The first major challenge in FER is the difference between controlled facial images and naturally occurring expressions. Laboratory datasets often contain frontal faces, consistent lighting, limited backgrounds, and expressions performed in response to instructions. Real-world images, in contrast, may contain partial occlusion, motion blur, varying camera distances, different head poses, and inconsistent illumination. These differences create a substantial gap between benchmark performance and practical reliability [1], [3], [5]. A second challenge concerns the ambiguity and subjectivity of expression labels. Two observers may interpret the same facial movement differently, particularly when expressions are subtle, mixed, or culturally dependent. Some datasets use predefined emotion categories, while others annotate facial action units, valence, or arousal. These annotation schemes describe different aspects of facial behavior and should not be treated as interchangeable targets [6]–[7].

Deep learning introduces an additional data-related challenge because complex neural networks generally require substantial training examples. Small datasets can encourage overfitting, especially when training and testing images share similar subjects or acquisition conditions. Imbalanced expression categories can also cause a model to favor frequently represented classes. Consequently, dataset composition, annotation quality, and evaluation design are as important as the architecture selected for FER [4], [8], [9].

## 3. DEEP-LEARNING ARCHITECTURES FOR FER

### 3.1. Convolutional neural networks

Convolutional neural networks (CNNs) are among the most widely used architectures for image-based FER. Their convolutional layers learn local visual patterns, while deeper layers combine these patterns into more abstract representations of facial structure and expression. Pooling and normalization operations can help manage spatial variation and improve computational efficiency. CNN-based approaches have demonstrated that learned representations can reduce dependence on manually designed facial descriptors [9], [10].

The effectiveness of a CNN depends on its architecture, training data, preprocessing, and optimization strategy. Face detection and alignment are commonly performed before classification to reduce irrelevant background information and standardize facial position. Data augmentation, including rotation, cropping, and brightness variation, can expose the network to a broader range of visual conditions. Lopes *et al.* [10] investigated CNN-based FER under limited training data and examined how preprocessing and sample presentation order influence learning performance.

### 3.2. Recurrent and temporal models

Static-image CNNs analyze individual frames, but facial expressions are dynamic processes that develop over time. Temporal information can distinguish a brief facial movement from a sustained expression and can capture transitions that are difficult to infer from a single image. Recurrent neural networks and long short-term memory architectures are designed to process sequences by maintaining information from earlier time steps. Their integration with CNN feature extractors provides a way to combine spatial facial information with temporal dynamics [11], [12].

Temporal modeling is particularly relevant to video-based FER and spontaneous-expression analysis. A sequence-based model can learn how facial features change during an expression rather than relying only on the final appearance. However, temporal approaches require consistent frame sampling, adequate video quality, and sufficient computational resources. They may also become sensitive to changes in frame rate, motion blur, and missing or poorly detected facial regions [11], [12].

### 3.3. Transfer learning, attention, and robust feature learning

Transfer learning provides an effective strategy when FER datasets are too small to train large networks from scratch. A model initially trained on a broader image dataset can supply general visual features that are subsequently adapted to facial-expression categories. Ramalingam and Garcia [13] examined transfer learning for FER and highlighted its relevance to training with limited expression data. This approach can reduce training requirements, although the transferred features must still be evaluated for their suitability to facial movements and target populations.

Attention mechanisms offer another way to improve recognition by emphasizing facial regions that contain useful expression information. The eyes, eyebrows, and mouth may contribute differently depending on the expression and image conditions. Zhang *et al.* [14] proposed a cycle-consistent adversarial attention-

transfer approach for FER in the wild, addressing the challenge of transferring useful attention information across domains. Such approaches are promising, but attention maps should not automatically be treated as complete explanations of a model's decisions. Deep learning can also incorporate structured facial information rather than treating the face as an undifferentiated image. Facial action units describe visible movements associated with specific facial muscle actions and can provide a more detailed representation than a small set of emotion labels. Investigated data-driven representations of facial activity and demonstrated their relevance to expression and micro-expression analysis. Joint use of action-unit information and expression categories may improve feature interpretation, although annotation quality remains essential.

## 4. DATASETS AND EVALUATION

Dataset selection strongly influences the conclusions that can be drawn from FER experiments. CK+ is a widely used benchmark containing posed facial expressions and action-unit annotations, making it useful for controlled comparisons [15]. DISFA provides spontaneous facial behavior with action-unit intensity annotations and supports research into more natural facial movements [16]. However, results obtained from these datasets should not be assumed to represent performance under all real-world conditions.

Larger datasets collected from unconstrained images have helped address some limitations of laboratory-based evaluation. RAF-DB contains facial images collected under varying poses and illumination, with annotations obtained through crowdsourcing and designed to represent more realistic expression variation [7]. AffectNet provides a large collection of facial images with both categorical expression labels and continuous valence-arousal annotations [6]. These resources broaden the range of evaluation conditions, although noisy labels, demographic representation, and class imbalance remain relevant concerns.

EmotioNet was developed to support large-scale facial action-unit and expression analysis through automatic annotation of images collected in the wild [7]. The Aff-Wild database and associated research extended evaluation toward spontaneous facial behavior in videos and continuous affect-related prediction [5]. FER2013 also became an influential benchmark for evaluating recognition under relatively unconstrained image conditions, including the challenges of representation learning and limited image resolution [8]. These datasets differ in their collection procedures, label definitions, and evaluation protocols, so direct comparison of reported accuracy requires caution.

Evaluation should therefore extend beyond a single accuracy value. Classification accuracy can be misleading when expression categories are imbalanced, while macro-averaged precision, recall, and F1-score provide additional information about class-specific performance. Cross-validation and subject-independent test partitions can reduce the risk of overly optimistic results caused by overlap between training and testing subjects. Cross-dataset evaluation is particularly valuable because it tests whether learned representations generalize to images collected under different conditions [1], [2], [4].

## 5. MULTIMODAL SYSTEMS AND APPLICATIONS

Camera-based FER remains attractive because ordinary cameras are widely available and can capture facial information without physical contact. Nevertheless, visual measurements can become unreliable when the face is partially hidden, poorly illuminated, or viewed from an extreme angle. Multimodal systems attempt to address these limitations by combining facial images with additional information, such as audio, depth, thermal imaging, or physiological signals. Samadiani *et al.* [17] reviewed these approaches and discussed how complementary sensors may improve robustness in challenging environments.

Deep-learning-based FER has potential applications in interactive systems, social robotics, behavioral research, and driver-monitoring technologies. In human–computer interaction, expression analysis may help interfaces respond to visible user behavior and support more adaptive interactions. In research settings, automated facial analysis can assist with large-scale observation of facial movements and action units. However, these applications require clear boundaries between detecting visible facial patterns and making broader psychological or medical judgments [3], [17].

## 6. LIMITATIONS, ETHICS, AND FUTURE RESEARCH

Despite considerable progress, generalization remains one of the most important unresolved challenges in FER. A model trained on one dataset may perform poorly when applied to different cameras, populations, lighting conditions, or cultural settings. Demographic imbalance can further affect error distribution across groups and lead to unequal system performance. Future studies should therefore include diverse training data, transparent demographic evaluation, and independent testing in realistic environments [1], [4], [6].

Ethical considerations are equally important because facial images are sensitive personal data. FER systems may be deployed in workplaces, public spaces, educational environments, or healthcare contexts

where individuals have limited control over image collection and interpretation. Facial expressions alone do not provide a definitive or universally valid measure of a person's thoughts, intentions, or emotional state. Responsible deployment should require informed consent where appropriate, privacy protection, clear disclosure of system limitations, and human oversight for consequential decisions [3], [17].

Further research should prioritize robust learning under domain shift, improved annotation practices, and interpretable evaluation. Models should be tested on spontaneous expressions and videos rather than relying exclusively on posed images. Multimodal approaches may provide useful complementary information, but their additional sensors introduce cost, synchronization, privacy, and integration challenges. The most meaningful progress will come from systems that combine technical performance with careful validation and responsible interpretation.

## 7. CONCLUSION

Deep learning has substantially advanced FER by enabling neural networks to learn discriminative facial representations from image and video data. CNNs provide effective spatial feature extraction, while recurrent models and temporal architectures can represent expression changes across video sequences. Transfer learning, attention mechanisms, and multimodal sensing offer additional strategies for improving recognition under limited data and challenging acquisition conditions. The growth of datasets such as RAF-DB, AffectNet, EmotioNet, and Aff-Wild has also expanded the scope of evaluation beyond controlled laboratory images. Nevertheless, high benchmark accuracy does not guarantee reliable performance in everyday environments. Dataset bias, ambiguous annotations, demographic imbalance, occlusion, illumination variation, and differences between posed and spontaneous expressions continue to limit generalization. Future FER systems should prioritize cross-dataset validation, fairness, explainability, privacy, and responsible application. Ultimately, FER should be understood as the analysis of observable facial behavior, not as an infallible method for determining a person's internal emotional state.

## REFERENCES

[1] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: A survey," *Symmetry*, vol. 11, no. 10, p. 1189, Sep. 2019, doi: 10.3390/sym11101189.

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

[3] S. Rajan, P. Chenniappan, S. Devaraj, and N. Madian, "Facial expression recognition techniques: A comprehensive survey," *IET Image Processing*, vol. 13, no. 7, pp. 1031–1040, May 2019, doi: 10.1049/iet-ipr.2018.6647.

[4] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, Jan. 2019, doi: 10.1109/TIP.2018.2868382.

[5] D. Kollias *et al.*, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, vol. 127, no. 6–7, pp. 907–929, 2019, doi: 10.1007/s11263-019-01158-4.

[6] 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, 2019, doi: 10.1109/TAFFC.2017.2740923.

[7] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5562–5570. doi: 10.1109/CVPR.2016.600.

[8] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Networks*, vol. 64, pp. 59–63, Apr. 2015, doi: 10.1016/j.neunet.2014.09.005.

[9] A. T. Lopes, E. Aguiar, A. F. 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, 2017, doi: 10.1016/j.patcog.2016.07.026.

[10] K. Shan, J. Guo, W. You, D. Lu, and R. Bie, "Automatic facial expression recognition based on a deep convolutional-neural-network structure," in *Proceedings - 2017 15th IEEE/ACIS International Conference on Software Engineering Research, Management and Applications, SERA 2017*, IEEE, Jun. 2017, pp. 123–128. doi: 10.1109/SERA.2017.7965717.

[11] M. Alam, L. S. Vidyaratne, and K. M. Iftekharuddin, "Sparse simultaneous recurrent deep learning for robust facial expression recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4905–4916, Oct. 2018, doi: 10.1109/TNNLS.2017.2776248.

[12] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Learning bases of activity for facial expression recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1965–1978, Apr. 2017, doi: 10.1109/TIP.2017.2662237.

[13] S. Ramalingam and F. Garzia, "Facial expression recognition using transfer learning," in *Proceedings - International Carnahan Conference on Security Technology*, IEEE, Oct. 2018, pp. 1–5. doi: 10.1109/CCST.2018.8585504.

[14] F. Zhang, T. Zhang, Q. Mao, L. Duan, and C. Xu, "Facial expression recognition in the wild: A cycle-consistent adversarial attention transfer approach," in *MM 2018 - Proceedings of the 2018 ACM Multimedia Conference*, New York, NY, USA: ACM, Oct. 2018, pp. 126–135. doi: 10.1145/3240508.3240574.

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

[16] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151–160, Apr. 2013, doi: 10.1109/T-AFFC.2013.4.

[17] N. Samadiani *et al.*, "A review on automatic facial expression recognition systems assisted by multimodal sensor data," *Sensors (Switzerland)*, vol. 19, no. 8, p. 1863, Apr. 2019, doi: 10.3390/s19081863.