Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Facial Expression Recognition Based on Deep Learning: A Review

Khalid Ibrahim Khalaf[1,2], Adnan Mohsin Abdulazeez [1]
khalid.ibrahim@auas.edu.krd, adnan.mohsin@dpu.krd.edu
[1] Technical Informatics, College of Akre, Duhok Polytechnic University, Duhok, Iraq
[2] Akre University for Applied Science, Duhok, Iraqffiliation

Article Information

Submitted : 22 Jan 2024
Reviewed: 27 Jan 2024
Accepted : 10 Feb 2024

Keywords

facial expression recognition, deep learning

Abstract

This review paper provides a comprehensive analysis of recent advancements in Facial Expression Recognition (FER) through various deep learning models. Seven state-of-the-art models are scrutinized, each offering unique contributions to the field. The MBCC-CNN model demonstrates improved recognition rates on diverse datasets, addressing the challenges of facial expression recognition through multiple branches and cross-connected convolutional neural networks. The Deep Graph Fusion model introduces a novel approach for predicting viewer expressions from videos, showcasing superior performance on the EEV database. Multimodal emotion recognition is explored in the EEG and facial expression fusion model, achieving high accuracy on the DEAP dataset. The Spark-based LDSP-TOP descriptor, coupled with a 1-D CNN and LSTM Autoencoder, excels in capturing temporal dynamics for facial expression understanding. Vision transformers for micro-expression recognition exhibit outstanding accuracy on datasets like CASMEI, CASME-II, and SAMM. Additionally, a hierarchical deep learning model is proposed for evaluating teaching states based on facial expressions. Lastly, a visionary transformer model achieves remarkable recognition accuracy of 100% on SAMM dataset, showcasing the potential of combining convolutional and transformer architectures. This review synthesizes key findings, highlights model performances, and outlines directions for future research in FER.

A. Introduction

Facial expression recognition (FER) has emerged as a pivotal area of research and application at the intersection of artificial intelligence (AI) and human-computer interaction. The ability to decipher and interpret human emotions through facial cues is a fundamental aspect of human communication, and the integration of technology to replicate this capability has garnered significant attention. This essay embarks on a comprehensive exploration of recent advancements in FER, encompassing diverse methodologies, novel architectures, and applications that extend beyond traditional emotion detection frameworks [1].

The landscape of FER has evolved dramatically, transitioning from traditional machine learning paradigms to the forefront of deep learning methodologies. This evolution is evident in the extensive exploration of convolutional neural networks (CNNs), which have demonstrated remarkable prowess in image-based tasks. One notable contribution is the Multiple Branch Cross-Connected CNN (MBCC-CNN), a novel architecture that integrates residual connections, Network in Network approaches, and a tree multibranch structure [2][3]. This synthesis enhances feature extraction and fuses information more effectively, showcasing improved recognition performance across benchmark datasets such as Fer2013, CKC, FERC, and RAF.

The pursuit of robust FER systems has led researchers to address inherent challenges, such as the class distribution mismatch problem. In response, innovative approaches like the Silhouette Coefficient-based Contrast Clustering algorithm have been proposed. This self-supervised learning method detects out-of-distribution data by examining intra-cluster and inter-cluster distances, demonstrating effectiveness in mitigating performance degradation caused by disparate data sources [4].

Real-world applications of FER extend beyond traditional contexts, with research delving into emotion recognition in varied scenarios. The Deep Graph Fusion model introduces a hybrid fusion approach, combining visual and auditory representations for predicting viewers' expressions from videos [5]. This nuanced model, equipped with graph convolutional networks and semantic embedding, outperforms conventional models, opening avenues for advancements in content creation, video recommendation, and the understanding of human emotional responses to visual stimuli.

AI's role in emotion care and well-being is underscored in the proposed emotion care system for autism disorder patients. Leveraging big data analysis, this system integrates EEG signals and facial expressions, creating a multimodal deep learning model for accurate emotion recognition. The exploration of multimodal approaches represents a frontier in FER, with experiments showcasing high accuracy rates on datasets like DEAP and MAHNOB-HCI [6].

The temporal dynamics of facial expressions in videos present another layer of complexity in FER research. Spark-based frameworks and innovative descriptors like the Local Directional Structural Pattern from Three Orthogonal Planes (LDSP-TOP) have been introduced to capture dynamic features effectively. The synthesis of 1-D CNNs and Long Short-Term Memory (LSTM) autoencoders enriches spatiotemporal feature learning, exemplifying the dedication to addressing challenges posed by video data [7].

Micro-expression recognition, a nuanced realm within FER, is tackled through a vision transformer-based model. This novel approach overcomes limitations of existing vision transformers by incorporating convolution patches, balancing local spatial relationships and global dependencies. The proposed model achieves exemplary performance on benchmark datasets such as CASME-I, CASME-II, and SAMM, positioning itself as a promising avenue for future research [8].

The exploration of facial expression recognition within educational contexts further solidifies its significance. The creation of the Intensity-based Facial Expression Dataset (EIDB-13) lays the foundation for evaluating teaching states objectively. The proposed hierarchical deep learning model, combining appearance and geometric features, showcases promising results, indicating its potential for fine-grained facial expression classification [9].

1.1. Motivation: The motivation behind undertaking this comprehensive review of 20 seminal works on Facial Expression Recognition (FER) lies in the transformative impact of deep learning on the field. With FER standing at the forefront of affective computing, the rapid evolution of deep learning methodologies has reshaped the landscape, addressing diverse challenges and opening new frontiers. The motivation is rooted in the recognition of FER's pervasive applications in technology, healthcare, and human-computer interaction. By critically examining and synthesizing these works, the goal is to contribute a nuanced understanding of the current state of FER, unveiling its potential societal impacts and guiding future research endeavors. This review seeks to motivate further exploration and innovation in FER, emphasizing its role in advancing technology and improving human-machine interactions.

1.2 Contribution: As the reviewer of the 20 selected works on Facial Expression Recognition (FER) based on deep learning, my contributions include critically evaluating each paper, synthesizing key insights and findings, categorizing and organizing the works, identifying technological trends, conducting in-depth analyses of methodologies, recognizing interdisciplinary dimensions, addressing real-world challenges, drawing insightful comparisons, evaluating contributions to the field, and offering guidance for future research directions. By providing a comprehensive and structured overview, this review aims to contribute to the understanding of the current state of FER, highlighting its technological advancements, societal implications, and potential avenues for further exploration.

B. Methodology

The exploration of facial expression recognition (FER) methodologies spans a diverse array of techniques, each tailored to address specific challenges and nuances within this intricate domain. The following section provides an in-depth insight into the methodologies employed in recent research, elucidating the key approaches and innovations that underpin the advancements in FER.

1. Architectural Innovations:

MBCC-CNN Model: Pioneering the realm of facial expression recognition, the Multiple Branch Cross-Connected CNN (MBCC-CNN) model amalgamates residual connections, Network in Network principles, and a tree multibranch structure. Constructed on the foundations of convolutional neural networks, this innovative architecture leverages the strengths of diverse methodologies to enhance feature extraction. The inclusion of shortcut cross connections fosters smooth data flow, addressing the challenge of insufficient feature extraction in individual branches. This model showcases superior recognition performance, as evidenced by competitive results across benchmark datasets [10].

2. Self-Supervised Learning for Class Distribution Mismatch:

Silhouette Coefficient-based Contrast Clustering: To mitigate the impact of class distribution mismatch, an ingenious self-supervised learning approach has been introduced. The Silhouette Coefficient-based Contrast Clustering algorithm discerns out-of-distribution data by scrutinizing intra-cluster and inter-cluster distances. This method rectifies the challenges posed by disparate data sources, ensuring that the model generalizes effectively across diverse datasets [11].

3. Real-world Scenario Applications:

Deep Graph Fusion Model: In the realm of real-world applications, the Deep Graph Fusion model takes center stage. This hybrid fusion model integrates visual and auditory modalities through graph convolutional networks and semantic embedding. The incorporation of a systematic approach to evaluate evoked expression scores from videos positions this model as a valuable tool for content creation and video recommendation. The use of the Evoked Expression from Videos (EEV) database substantiates its prowess in outperforming conventional models [12].

4. Multimodal Emotion Recognition:

EEG and Facial Expression Fusion Model: Multimodal emotion recognition takes precedence with the fusion of electroencephalogram (EEG) signals and facial expressions. This end-to-end model employs a pre-trained convolutional neural network for facial feature extraction and introduces an attention mechanism for enhanced feature selection. The spatial features extracted from EEG signals are integrated, demonstrating exceptional classification accuracy on datasets such as DEAP and MAHNOB-HCI [13].

5. Temporal Dynamics in Video Data:

Spark-based LDSP-TOP Descriptor: Addressing the challenges posed by the temporal dynamics of facial expressions in videos, a Spark-based framework is harnessed. The Local Directional Structural Pattern from Three Orthogonal Planes (LDSP-TOP) descriptor emerges as a dynamic feature descriptor [14]. Augmented by a 1-D convolutional neural network (CNN) and Long Short-Term Memory (LSTM) autoencoder, this methodology effectively captures spatiotemporal features, enhancing recognition rates on datasets like CKC, Oulu-CASIA, and LIRIS-CSE[15].

6. Micro-Expression Recognition with Vision Transformer:

Vision Transformer Based on Convolution Patches: Delving into the nuances of micro-expression recognition, a vision transformer model based on convolution patches is proposed. Overcoming the limitations of existing vision transformers, this novel approach balances local spatial relationships and global dependencies. Achieving exemplary performance on datasets such as CASME-I, CASME-II, and SAMM, this methodology integrates convolutional layers and transformers for improved spatial information and feature extraction [16][17].

7. Educational Contexts and Teaching Quality Evaluation:

Hierarchical Deep Learning Model: Within educational contexts, the methodology revolves around the creation of the Intensity-based Facial Expression Dataset (EIDB-13) and a hierarchical deep learning model[18]. This model combines appearance and geometric features, providing a fine-grained classification of facial expressions. The use of SoftMax function and the consideration of Top-2 prediction results showcase an effective approach for evaluating teaching states objectively[19].

In summation, the methodologies encapsulated in recent FER research underscore the dynamic and multifaceted nature of this field. From architectural innovations to addressing real-world challenges and venturing into multimodal and temporal domains, each methodology contributes to the collective progression of facial expression recognition, promising enhanced accuracy, and applicability across diverse scenarios.

C. Literature Review

Facial Expression Recognition (FER) has witnessed substantial progress driven by the integration of deep learning techniques in recent years. The reviewed articles collectively contribute to expanding the horizons of FER methodologies [20]. The MBCC-CNN model brings innovation through a multi-branch, cross-connected convolutional neural network, addressing challenges in feature extraction and recognition rates. The Deep Graph Fusion model explores novel avenues by predicting viewer expressions from videos, tapping into both visual and auditory modalities. The EEG and Facial Expression Fusion model showcases the potential of multimodal emotion recognition, advancing the understanding of emotional states through the fusion of electroencephalogram signals and facial expressions. Additionally, the Spark-based LDSP-TOP descriptor, coupled with a 1-D CNN and LSTM Autoencoder, excels in capturing temporal dynamics in facial expressions [21]. Vision transformers stand out in micro-expression recognition, emphasizing the importance of combining local and global information for accurate classification. The proposed hierarchical deep learning model introduces a unique scheme for evaluating teaching states based on facial expressions, providing a foundation for objective assessment of teaching quality [22]. Furthermore, the visionary transformer model achieves exceptional accuracy, underscoring the promising trajectory of combining convolutional and transformer architectures in FER. This literature review encapsulates the diverse landscape of recent developments in FER, offering insights into novel methodologies and paving the way for future exploration in this dynamic field.

Dina et al. [23] Addresses the challenge faced by visually impaired individuals in perceiving non-verbal cues, hindering their conversational interactions and daily activities. Leveraging advancements in computer vision, the authors propose a partial transfer learning approach using a custom-trained Convolutional Neural Network (CNN) for automated facial emotion recognition. The model achieves a notable recognition accuracy of 82.1% on the enhanced Facial Expression Recognition 2013

(FER2013) dataset and operates efficiently on edge devices with limited resources. Specifically designed for visually impaired individuals, the system overcomes limitations of existing facial expression recognition systems, providing a portable, lightweight, and accurate assistive solution. The model successfully recognizes happy, sad, and surprise emotions with high accuracy, though it shows a lower accuracy for anger, disgust, and fear. The authors suggest future improvements by training the model on more samples from these emotions and plan to implement the system on Raspberry Pi for practical testing and enhancement.

Lee and Yoo. [24] Focuses on enhancing facial expression recognition (FER) performance in computer vision, acknowledging the challenges posed by varying degrees of similarity among different expressions. The authors propose a novel divide-and-conquer learning strategy, utilizing MobileNet for face detection and ResNet-18 as the backbone deep neural network. The approach involves categorizing similar facial expressions based on confusion matrix analysis of the trained model's inference results, followed by retraining the model with these categorized groups. Evaluation on thermal (Tufts and RWTH) and RGB (RAF and FER2013) datasets reveals notable improvements, achieving accuracies of 97.75%, 86.11%, 90.81%, and 77.83%, respectively. The study introduces an efficient CNN learning strategy, emphasizing the need for an integrated architecture to simultaneously address subproblems and reduce computational costs. Future directions include extending the proposed method to Vision Transformer-based models and its application in diverse recognition tasks beyond facial expressions, such as voice, action, object, and text recognition.

Pu and Nie. [25] Aims to enable robots to comprehend human emotions through a human-computer interaction system, focusing on facial expressions. A facial expression recognition (FER) model based on convolutional neural networks (CNN) and channel attention is constructed. To deploy this model on portable devices, a depth-separable convolution filter pruning algorithm, utilizing principal component analysis (PCA), is proposed. The FER algorithm achieves a remarkable 99% recognition accuracy with an average of 80.39%, using fewer parameters compared to other algorithms. The lightweight network, employing the pruning strategy, demonstrates a 93.24% correct rate for facial expression recognition with a 40% pruning rate. The proposed algorithm significantly accelerates the model, reducing memory occupancy by 41% and improving classification accuracy. The study emphasizes the use of deep separable convolution and PCA techniques to enhance the ResNet network structure's convolutional layer. The resulting model achieves high recognition accuracy and efficiency, but the recognition speed needs further optimization for ideal performance. Additionally, the study suggests exploring parallel pruning for further optimization of multi-layer models.

Zhou et al. [26] Addresses the limitations of intelligent facial recognition systems, focusing on the challenges of limited biometric features compared to methods like iris and fingerprint authentication. The proposed solution involves extracting facial feature information through facial expression key points and employing a spatiotemporal graph convolutional network (STGCN) fused with an attention mechanism for organizing and matching feature data. The improved facial recognition system exhibits an accuracy of 89% in different datasets, demonstrating higher accuracy and operational efficiency compared to existing methods. The study further proposes an enhanced STGCN-SA facial recognition system, incorporating graph convolutional neural networks (GCNNs) and attention mechanisms to capture spatiotemporal sequences of facial expression key points. Experimental results show significant improvements in facial recognition accuracy across various datasets, exceeding 89%. The technology is deemed suitable for high-security applications such as school access control and bank security systems. However, the study acknowledges challenges in recognizing side facial features and suggests future research directions to enhance performance in various orientations, potentially incorporating additional information to improve side facial feature recognition.

Hussain et al. [27] Introduced an automatic emotion recognition (AER) system using deep learning with electroencephalogram (EEG) signals, addressing the challenges of traditional methods relying on hand-engineered features. The proposed system employs a lightweight pyramidal one-dimensional convolutional neural network (LP-1D-CNN) and a two-level ensemble classifier for accurate emotion prediction. The study validates the method using the DEAP dataset, analyzing EEG signals from five brain regions. Results reveal that the frontal brain region plays a dominant role, achieving high accuracies of 98.43% and 97.65% for distinguishing high valence vs. low valence and high arousal vs. low arousal states. The LP-1D-CNN model, designed with a pyramid architecture, surpasses traditional approaches, showcasing the superiority of deep learning in AER from EEG signals. The proposed system's accuracy and potential applications in mental health, including obsessive-compulsive disorder, highlight its effectiveness in emotion recognition.

Lan et al. [28] Addressed challenges in facial expression recognition arising from natural environment noise, illumination, and occlusion. The proposed Multi-Region Coordinate Attentional Residual (MrCAR) model aims to enhance recognition accuracy. The model comprises three main components: 1) Multi-region input: Face detection and alignment using MTCNN results in cropped eye and mouth regions, allowing for better local and global feature extraction, minimizing environmental noise impact. 2) Feature extraction module: A residual element, combined with Coordinate Attention (CA-Net) and multi-scale convolution, forms a coordinate residual attention module, enhancing the model's ability to distinguish subtle expression changes and optimize key feature utilization. 3) Classifier: Arcface Loss is employed to simultaneously enhance intra-class tightness and inter-class differences, reducing misclassification of negative expressions. Experimental results on CK+, JAFFE, FER2013, and RAF-DB datasets demonstrate accuracy rates of 98.78%, 99.09%, 74.50%, and 88.26%, respectively. MrCAR outperforms advanced models in expression classification tasks, particularly in natural scenarios. Future work will focus on real-time video facial expression recognition.

He Y [29] Present a novel approach to facial expression recognition by introducing a Multi-Branch Attention Convolutional Neural Network (CNN). The model employs a multi-branch structure for feature extraction from facial images, merging features from three branches. To enhance feature extraction and recognition performance, a Convolutional Block Attention Module is integrated as an attention mechanism. Additionally, the model optimizes parameters and reduces computational load through the use of depth-wise separable convolutions. Experimental results on the FER2013, FERPLUS, and CKC datasets demonstrate recognition rates of 69.49%, 84.633%, and 99.39%, respectively. The proposed method exhibits higher efficiency in feature extraction compared to traditional deep learning models, achieving high accuracy without complex artificial feature engineering. The Multi-Branch Attention CNN, incorporating attention modules and depth-wise separable convolutions, outperforms classic classification models, showcasing its potential for improved expression recognition and generalization in real video data.

Shahzad et al. [30] Addresses the challenge of facial expression recognition (FER) in the context of widespread mask-wearing during the COVID-19 pandemic. Recognizing the limitations of traditional unimodal techniques, the study proposes a novel multimodal approach that integrates facial and vocal expressions using deep learning. Two standard datasets, M-LFW-F and CREMA-D, are employed to capture facial and vocal emotional expressions, respectively. The multimodal neural network, incorporating fusion techniques, outperforms unimodal methods in FER, achieving an accuracy of 79.81%, significantly surpassing the 68.81% accuracy of unimodal techniques. The proposed approach demonstrates superior performance in recognizing facial expressions under masked conditions. The paper emphasizes the importance of addressing challenges such as overfitting through appropriate regularization and fusion techniques, leading to a notable improvement in accuracy. However, the study acknowledges limitations in various orientation scenarios and underscores the need for careful consideration of different modalities with distinct scales and units. Additionally, the paper highlights the potential biases in training and assessment datasets, emphasizing the importance of cultural diversity for more robust models. The study encourages future research to incorporate a broader range of expressions to enhance the inclusivity and reliability of facial expression analysis, especially in diverse cultural contexts.

Perveen et al. [31] Presents a novel approach for facial micro-expression recognition (MER) utilizing a multi-stream deep convolutional neural network with ensemble classification. The multi-stream network incorporates features from a residual network, densely connected convolutional network, and visual geometry group. To manage the resource-intensive high-dimensional features from these architectures, principal component analysis (PCA) is applied for dimensionality reduction. The ensemble classification technique, stacking, is employed with three base learners (random tree, J48, random forest) and a meta learner (random forest). Experimental evaluations on CASME-II, CASME2, SMIC, and SAMM datasets demonstrate that the proposed approach outperforms twelve existing methods in terms of accuracy and time efficiency, achieving impressive results of 98.68%, 99.39%, 96.01%, and 99.80% on SMIC, CASME-II, CAS(ME)2, and SAMM datasets, respectively. The features extracted from deep architectures (DenseNet-121, VGG-16, and ResNet-50) are significantly reduced in dimensionality through PCA, showcasing the effectiveness of the proposed approach. Additionally, the ensemble technique (stacking) in the proposed approach surpasses other ensemble techniques and classifiers, highlighting its contribution to improved MER. Future work is planned to extend the model to challenging environments and explore practical applications, including integration with IoT devices, aiming for compact deep learning solutions with enhanced accuracy.

Fang et al. [32] Tackles the challenge of class distribution mismatch in self-supervised learning for semi-supervised facial expression recognition when using facial expression data from a large face recognition database as unlabeled data. The authors propose a silhouette coefficient-based contrast clustering algorithm to assess the separation between clusters, effectively detecting out-of-distribution data. Additionally, they introduce a pseudo-labeling rethinking strategy that aligns soft pseudo-labels from a fine-tuned network with the contrast clustering results to produce reliable pseudo-labels. Experiments on three in-the-wild datasets (RAF-DB, FERPlus, and AffectNet) demonstrate the effectiveness of the proposed method, showcasing its superiority over state-of-the-art approaches. The authors acknowledge the class imbalance problem in the training dataset and express their intent to address this issue in future work, considering its potential impact on the recognition accuracy of expressions with smaller sample sizes.

Cuiping et al. [33] Introduces a novel facial expression recognition method based on a multiple branch cross-connected convolutional neural network (MBCC-CNN). The proposed method, in contrast to traditional machine learning approaches, demonstrates enhanced effectiveness in feature extraction. The MBCC-CNN model is constructed by integrating residual connection, Network in Network, and tree structure approaches, incorporating a shortcut cross connection for smoother data flow between networks, thereby improving the feature extraction capabilities of each receptive field. The method effectively fuses features from each branch, addressing the challenge of insufficient feature extraction and enhancing recognition performance. Experimental results on Fer2013, CKC, FERC, and RAF datasets indicate recognition rates of 71.52%, 98.48%, 88.10%, and 87.34%, respectively. Comparative analysis with recent methods demonstrates the superior facial expression recognition performance and robustness of the proposed MBCC-CNN model. The paper also presents an expression recognition system using the MBCC-CNN method, showcasing its quick and accurate recognition capabilities for potential applications in intelligent and real-time environments. The authors express their intent to explore facial expression recognition in complex environments in future work.

Pinto et al. [34] Presents a Systematic Literature Review (SLR) that explores techniques and algorithms for facial expression recognition, with a focus on convolutional neural network (CNN) models in the context of deep learning. The review analyzes studies related to expression and micro-expression recognition using CNNs, particularly emphasizing the performance of models on databases with laboratory-controlled images.

The results of the SLR indicate that CNN models, such as VGG and ResNet, have demonstrated excellent performance in facial expression recognition tasks, especially when applied to databases with controlled laboratory images. The review highlights the importance of the dataset used, with an emphasis on laboratory-controlled images without external interference.

Furthermore, the paper compares CNN models, including VGG and ResNet, and discusses their suitability for different scenarios. VGG, with its simpler network structure and relatively fewer parameters, may be preferred in situations where simplicity and ease of implementation are crucial. On the other hand, ResNet, with its deeper and more complex architecture, excels in tasks requiring the learning of complex and abstract features, making it suitable for large and complex datasets.

The SLR also mentions other notable algorithms, such as YOLOv3 for object detection and VGG-16 for image classification tasks. VGG-16, known for its effectiveness in image classification, is highlighted for its simplicity and capability of achieving high accuracy on various benchmark datasets.

While the paper acknowledges some limitations, such as the restriction of the search period and the potential exclusion of short articles with original ideas, it emphasizes the importance of the obtained data for future research. The authors suggest that new publications continue to contribute to the field of facial expression recognition, introducing new CNN models and databases, and fostering innovative approaches to recognizing expressions and faces.

As future work, the authors plan to update the SLR with new databases and incorporate additional search terms. The research is part of a broader project aiming to develop a tool for emotion detection in scenarios of violence against women.

Liu et al. [35] Proposed an end-to-end deep model to enhance facial expression recognition accuracy by addressing challenges related to the complex environment and diverse emotion expressions. The key focus is on improving facial feature extraction, as more discriminative features are deemed essential for accurate facial expression recognition.

The proposed model comprises three main stages: data pre-processing, feature extraction, and classification. The authors emphasize the importance of data pre-processing, introducing a data enhancement method to locate the face target and enhance image contrast. Following this, a hybrid feature representation method is introduced, combining four typical feature extraction techniques to obtain more discriminative features. Finally, an effective deep model is designed for training and testing samples, aiming to achieve optimal parameters with reduced computational cost. The effectiveness of the proposed hybrid feature representation method is demonstrated through ablation study results, showing improved recognition accuracy. The model is comprehensively evaluated through experiments on three benchmark datasets: FER2013, AR dataset, and CKC dataset. The proposed model achieves recognition rates of 94.5%, 98.6%, and 97.2%, respectively.

As part of future work, the authors express the intention to focus on face alignment technology and propose a new deep model to enhance recognition rates, particularly under conditions of large-area occlusion of the face.

Ho et al. [36] Introduced a novel hybrid fusion model, termed "deep graph fusion," designed to predict viewers' elicited expressions from videos through the amalgamation of visual and audio representations. The proposed system unfolds in four stages: firstly, features are extracted for visual and auditory modalities using CNN-based pre-trained models. Subsequently, these features are reconstructed into graph outlines, and graph convolutional networks (GCNs) are applied for node embedding. The fusion modules are then employed to combine the graph representations from the visual and auditory branches. Finally, the fused features are utilized to predict evoked scores for emotional classes using Sigmoid activation, and a semantic embedding loss is introduced for improved semantic understanding of textual emotions. Evaluation on the Evoked Expression from Videos (EEV) database reveals superior performance, especially in terms of the Pearson correlation coefficient, surpassing traditional models. The authors outline future refinements, including enhancements to the training approach considering inter-segment connections, and they acknowledge current limitations related to predicting evoked expressions for individual movie sections due to memory device constraints, intending to address these aspects in their subsequent work.

Wang et al. [37] Presented a comprehensive emotion care system tailored for autism disorder patient training, emphasizing big data analysis and facial expression detection. The system seamlessly integrates both camera and Internet of Things (IoT)-enabled devices to capture emotional cues. Key components of the system encompass a sophisticated data processing technique, leveraging deep learning with a convolutional neural network (CNN) model based on the MobileNet V1 structure for real-time and offline facial expression recognition. The CNN model is adeptly trained using two emotional datasets, namely the FER-2013 dataset and a novel dataset introduced in the paper - MCFER, collected at the University of Ottawa. Employing a two-stage strategy and joint supervision, the authors demonstrate enhanced model performance, validated on datasets CKC and MCFER. The system achieves an impressive accuracy of 95.89% in recognizing six facial expressions, boasting the capability to detect and track multiple faces in both real-time and offline scenarios. Further, an Android application is developed, featuring saving and free modes, with the latter operating at a commendable speed of up to 12 frames per second. The authors, considering resource constraints in embedded systems, suggest future improvements, emphasizing model quantification to optimize memory space and enhance processing speed. This emotion care system, driven by deep learning and facial expression recognition, emerges as a promising tool for autism disorder patient training, showcasing commendable accuracy and efficient real-time processing on mobile devices.

Uddin et al. [38] Introduced an innovative framework for dynamic facial expression recognition, with a particular emphasis on the extraction and modeling of temporal dynamics from videos. Their approach, implemented on the Apache Spark distributed computing environment, unfolds through several pivotal stages. Firstly, a dynamic feature descriptor, the Local Directional Structural Pattern from Three Orthogonal Planes (LDSP-TOP), is introduced to analyze the structural aspects of local dynamic texture, ensuring a stable representation of facial dynamics. Subsequently, a 1-D Convolutional Neural Network (CNN) with residual connections is devised to capture additional discriminative features from the dynamic facial expressions. The inclusion of a Long Short-Term Memory (LSTM) autoencoder enhances the model's ability to grasp spatiotemporal features and comprehend temporal variations in facial expressions. Leveraging Apache Spark for distributed computing, the proposed framework is intended to boost processing efficiency. The authors conduct extensive experiments, evaluating the framework's performance and scalability on datasets like CKC, Oulu-CASIA, and LIRIS-CSE. While acknowledging the absence of facial landmark information in this study, the authors express a commitment to exploring the integration of facial landmark features in future work. Furthermore, they outline plans to extend the application of their approach to diverse video analysis domains such as video retrieval and action recognition. This comprehensive framework addresses challenges in facial expression recognition from video data, amalgamating innovative feature descriptors with deep learning techniques and highlighting scalability through Apache Spark.

Indolia et al. [39] Tackle the challenges associated with recognizing micro-expressions, subtle facial muscle changes that are often involuntary and easily concealed. They present a novel vision transformer based on convolution patches, aiming to overcome limitations in existing vision transformers regarding capturing local spatial relationships in images. The proposed algorithm generates feature maps using convolutional filters and applies them to a transformer model as fixed-size image patches for classification, leveraging both convolutional layers and transformers to capture spatial information and global dependencies. The contributions include addressing issues with CNNs, such as overfitting, and the constraints of existing vision transformers, specifically the loss of local correlation due to fixed image patches. Implementation considerations involve utilizing high-performance computational resources (Nvidia A100) and mitigating overfitting through layer normalization and dropout mechanisms. Performance evaluation on benchmark datasets (CASME-I, CASME-II, and SAMM) demonstrates significant improvements, with classification accuracy reaching 95.97%, 98.59%, and 100%, respectively. The study acknowledges challenges, including unbalanced datasets, and proposes future work to address this through data augmentation techniques. Furthermore, the exploration of in-the-wild datasets for real-life applications and the suggestion of deep continual learning to identify unknown emotion categories beyond predefined classes highlight the comprehensive nature of the proposed vision transformer in micro-expression recognition.

Wang et al. [40] Introduced a multimodal emotion recognition model that combines electroencephalogram (EEG) signals and facial expressions to enhance classification accuracy. The approach integrates information from both modalities, utilizing a pre-trained Convolutional Neural Network (CNN) for facial feature extraction, complemented by an attention mechanism to enhance frame-specific feature extraction. EEG signal processing involves applying CNNs to extract spatial features, employing local and global convolution kernels for comprehensive feature learning. Feature-level fusion combines the extracted features from facial expressions and EEG signals, feeding them into a classifier for emotion recognition, specifically predicting valence and arousal labels. Performance evaluation on the DEAP and MAHNOB-HCI datasets demonstrates high accuracy, with results indicating 96.63% and 97.15% accuracy for valence and arousal on the DEAP dataset, and 96.69% and 96.26% accuracy on the MAHNOB-HCI dataset, respectively. The study concludes that the proposed model effectively recognizes emotions, and the fusion of modalities surpasses the use of EEG or facial expressions alone. Future work is suggested to explore more reliable pre-training models for facial expression features and incorporate additional modalities, such as non-physiological signals, for a more enriched multimodal emotion recognition model. The study contributes significantly to the field of emotion recognition, showcasing promising results and paving the way for further advancements.

Kim et al. [41] Introduced a hierarchical deep learning-based facial expression recognition (FER) system that emphasizes the fusion of appearance and geometric features. The approach employs a hierarchical deep learning scheme, incorporating an appearance feature-based network that extracts holistic features using preprocessed Local Binary Pattern (LBP) images. These LBP features contain Action Units (AUs) information related to facial muscle movements during expressions. Additionally, a geometric feature-based network captures dynamic features by learning coordinate changes of AUs landmarks, focusing on muscle movement during facial expressions. The system combines SoftMax function results from both appearance and geometric features in a hierarchical structure to mitigate errors associated with the second-highest emotion prediction (Top-2). Furthermore, the paper proposes a technique to generate facial images with a neutral emotion using the autoencoder technique, allowing for the extraction of dynamic facial features without sequence data. Evaluation on CKC and JAFFE datasets, standard datasets in FER, demonstrates impressive results, achieving 96.46% accuracy on CKC and 91.27% accuracy on JAFFE. The proposed algorithm outperforms recent alternatives, showing up to 3% accuracy improvement and 1.3% average improvement on CKC and up to 7% accuracy enhancement and 1.5% average improvement on JAFFE. The study significantly contributes to FER systems by effectively combining static appearance and dynamic geometric features within a hierarchical deep learning framework, showcasing substantial accuracy improvements across diverse datasets.

Zheng et al. [42] Addressed the challenge of objectively evaluating teaching quality by proposing the Intensity-based Facial Expression Dataset (EIDB-13), which focuses on 13 types of facial expressions with an emphasis on intensity. The dataset, comprising 10,393 facial images from various sources, aims to enable fine-grained analysis of teachers' expressions in real classroom settings. The authors introduce a facial expression recognition algorithm that combines Convolutional Neural Network (CNN) and an attention mechanism, specifically integrating the Convolutional Block Attention Module (CBAM) into the InceptionResNetV2 model. Migration learning, utilizing InceptionResNetV2 as the migration network, is employed to address overfitting issues with small sample datasets. The proposed algorithm achieves a notable 78% classification accuracy on the EIDB-13 dataset, emphasizing intensity-based facial expressions, and an 88% accuracy on the RAF-DB dataset. The recognition algorithm is applied to assess teaching quality by detecting the frequency and intensity of teachers' facial expressions in classrooms, aiming to provide an objective reference for teaching assessment and analyze students' interest. The system, integrated with the MTCNN face detection algorithm, exhibits real-time face detection and expression recognition. The paper acknowledges challenges in fine-grained emotional expression segmentation and emphasizes ongoing efforts to enhance deep learning robustness for feature extraction and address small sample challenges in large-scale intelligent recognition based on facial expressions.

Table 1. Selected Papers

Paper Number	Title and Focus	Key Contributions	Performance Metrics	Challenges and Future Directions
[23], 2023	Assistive Facial Emotion Recognition for Visually Impaired	Partial transfer learning for facial emotion recognition for visually impaired individuals	82.1% accuracy on FER2013	Improve accuracy for anger, disgust, and fear; Implement on Raspberry Pi for practical testing
[24], 2023	Divide-and-Conquer Learning for Facial Expression Recognition	Novel learning strategy for FER using MobileNet and ResNet-18	97.75%, 86.11%, 90.81%, 77.83% on thermal and RGB datasets	Extend to Vision Transformer models; Apply to various recognition tasks beyond facial expressions
[25], 2023	Lightweight Facial Expression Recognition for Human-Robot Interaction	FER model with depth-separable convolution pruning for portable devices	99% accuracy with 40% pruning rate	Optimize recognition speed; Explore parallel pruning for multi-layer models
[26], 2023	Enhanced Facial Recognition with STGCN and Attention Mechanism	STGCN fused with attention mechanism for improved facial recognition	89% accuracy	Improve side facial feature recognition; Explore additional information for

				orientation variations
[27], 2023	Deep Learning for Emotion Recognition from EEG Signals	LP-1D-CNN and ensemble classifier for AER from EEG signals	98.43%, 97.65% accuracy for high vs. low valence and arousal	Highlight effectiveness in mental health; Explore applications in disorders like OCD
[28], 2023	MrCAR Model for Facial Expression Recognition in Natural Environments	Multi-region, Coordinate Attentional Residual model for accurate expression recognition	98.78%, 99.09%, 74.50%, 88.26% accuracy on CK+, JAFFE, FER2013, RAF-DB datasets	Focus on real-time video expression recognition
[29], 2023	Multi-Branch Attention CNN for Facial Expression Recognition	Multi-branch structure with attention modules for efficient feature extraction	69.49%, 84.633%, 99.39% accuracy on FER2013, FERPLUS, CKC datasets	Highlight efficiency without complex feature engineering
[30], 2023	Multimodal Facial Expression Recognition under Mask Wearing	Multimodal approach integrating facial and vocal expressions for FER	79.81% accuracy	Address overfitting; Consider modality-specific challenges; Enhance inclusivity and reliability
[31], 2023	Ensemble Classification for Facial Micro-Expression Recognition	Multi-stream network with ensemble classification for MER	98.68%, 99.39%, 96.01%, 99.80% accuracy on SMIC, CASME-II, CAS(ME)2, SAMM datasets	Extend to challenging environments; Explore IoT integration
[32], 2023	Self-Supervised Facial Expression Recognition with Contrast Clustering	Silhouette coefficient-based clustering algorithm for semi-supervised FER	Superior performance on RAF-DB, FERPlus, AffectNet datasets	Address class imbalance; Explore expression categories with smaller sample sizes
[33], 2021	Multiple Branch Cross-Connected CNN for Facial Expression Recognition	MBCC-CNN for improved feature extraction in FER	71.52%, 98.48%, 88.10%, 87.34% accuracy on Fer2013, CKC, FERC, RAF datasets	Explore FER in complex environments
[34], 2023	Systematic Literature Review on CNN Models for Facial Expression Recognition	Review of CNN models for expression recognition on controlled datasets	Comparison of VGG and ResNet; Emphasis on dataset quality	Update with new databases and search terms

[35], 2021	Hybrid Feature Representation for Facial Expression Recognition	Hybrid feature representation method for improved FER	94.5%, 98.6%, 97.2% accuracy on FER2013, AR, CKC datasets	Focus on face alignment technology; Enhance recognition rates
[36], 2021	Deep Graph Fusion for Video-Based Emotion Prediction	Fusion of visual and audio representations using deep graph fusion	Superior performance on EEV database	Enhance training approach; Address memory device constraints
[37], , 2021	Emotion Care System for Autism Disorder Patients	IoT-enabled emotion care system using CNN for facial expression recognition	95.89% accuracy on CKC and MCFER datasets	Optimize memory space for embedded systems; Explore real-time applications
[38], 2021	Apache Spark-Based Dynamic Facial Expression Recognition	LDSP-TOP, CNN, and LSTM for dynamic FER on Apache Spark	Significant improvements on CKC, Oulu-CASIA, LIRIS-CSE datasets	Explore facial landmark integration; Extend to diverse video analysis domains
[39], 2023	Vision Transformer for Micro-Expression Recognition	Vision transformer for micro-expression recognition with improved spatial relationships	95.97%, 98.59%, 100% accuracy on CASME-I, CASME-II, SAMM datasets	Address unbalanced datasets; Explore in-the-wild datasets; Deep continual learning
[40], 2023	Multimodal Emotion Recognition with EEG and Facial Expressions	Multimodal model combining EEG and facial expressions for emotion recognition	96.63%, 97.15%, 96.69%, 96.26% accuracy on DEAP and MAHNOB-HCI datasets	Explore reliable pre-training models; Introduce additional modalities
[41], 2019	Hierarchical Deep Learning for Facial Expression Recognition	Hierarchical deep learning for FER with fusion of appearance and geometric features	96.46%, 91.27% accuracy on CKC and JAFFE datasets	Improve recognition speed; Explore real-world applications
[42], 2020	Intensity-Based Facial Expression Dataset for Teaching Quality Evaluation	EIDB-13 dataset and recognition algorithm for teaching quality assessment	78% accuracy on EIDB-13; 88% on RAF-DB	Improve deep learning robustness; Address small sample challenges

D. Discussion

The discussions below provide an overview of the key insights derived from the summaries of the twenty papers on facial expression recognition. Each discussion point corresponds to a specific paper, identified by its number.

a. A Lightweight Facial Emotion Recognition System Using Partial Transfer Learning for Visually Impaired People

The proposed partial transfer learning approach addresses the challenges faced by visually impaired individuals, providing a notable recognition accuracy of 82.1% on the FER2013 dataset. The focus on custom-trained CNNs for automated facial emotion recognition contributes to a portable, lightweight, and accurate assistive solution[23].

b. CNN Learning Strategy for Recognizing Facial Expressions

The novel divide-and-conquer learning strategy, utilizing MobileNet and ResNet-18, showcases significant improvements in facial expression recognition accuracy across thermal and RGB datasets. The emphasis on integrated architectures and the potential extension to Vision Transformer-based models reflects a comprehensive approach[24].

c. Convolutional Channel Attentional Facial Expression Recognition Network and Its Application in Human–Computer Interaction:

This study introduces a lightweight facial expression recognition model suitable for human-robot interaction, achieving remarkable recognition accuracy (99%) with a 40% pruning rate. Future directions include optimizing recognition speed and exploring parallel pruning techniques for multi-layer models[25].

d. Design of an Intelligent Laboratory Facial Recognition System Based on Expression Keypoint Extraction:

Addressing limitations in facial recognition systems, the proposed STGCN with an attention mechanism significantly improves accuracy to 89%. Challenges related to side facial features recognition are acknowledged, emphasizing the need for future research to enhance performance in various orientations[26].

e. Emotion Recognition System Based on Two-Level Ensemble of Deep-Convolutional Neural Network Models:

Introducing an automatic emotion recognition system using deep learning with EEG signals, this paper presents a lightweight pyramidal CNN achieving high accuracies (98.43%, 97.65%). The study underscores the effectiveness of deep learning in AER from EEG signals, showcasing potential applications in mental health[27].

f. Expression Recognition Based on Multi-Regional Coordinate Attention Residuals:

The MrCAR model addresses challenges in facial expression recognition from natural environments, achieving impressive accuracy rates across diverse datasets. Future work is planned for real-time video facial expression recognition, suggesting a focus on practical applications[28].

g. Facial Expression Recognition Using Multi-Branch Attention Convolutional Neural Network:

The Multi-Branch Attention CNN introduces a novel approach with a multi-branch structure and attention modules, demonstrating high accuracy rates on various datasets. The emphasis on efficiency without complex feature engineering highlights its potential for real-world applications[29].

h. Multi-Modal CNN Features Fusion for Emotion Recognition:

In the context of mask-wearing during the COVID-19 pandemic, the study proposes a multimodal approach integrating facial and vocal expressions, outperforming unimodal techniques. Challenges such as overfitting and modality-specific considerations are acknowledged, emphasizing the importance of cultural diversity in datasets[30].

i. Multi-Stream Deep Convolution Neural Network With Ensemble Learning for Facial Micro-Expression Recognition:

The paper introduces a multi-stream network with ensemble classification for facial micro-expression recognition, showcasing impressive accuracy rates. Future work aims to extend the model to challenging environments and explore integration with IoT devices, emphasizing compact deep learning solutions[31].

j. Rethinking Pseudo-Labeling for Semi-Supervised Facial Expression Recognition With Contrastive Self-Supervised Learning:

Addressing class distribution mismatch, the paper proposes a silhouette coefficient-based contrast clustering algorithm for self-supervised facial expression recognition. The method demonstrates superiority over existing approaches, with future work focused on addressing class imbalance and exploring expressions with smaller sample sizes[32].

k. A Facial Expression Recognition Method Based on a Multibranch Cross-Connection Convolutional Neural Network:

The Multiple Branch Cross-Connected CNN introduces an effective method for feature extraction, leading to enhanced facial expression recognition performance. The model's quick and accurate capabilities suggest potential applications in intelligent and real-time environments[33].

l. A Systematic Review of Facial Expression Detection Methods:

The systematic literature review emphasizes the excellent performance of CNN models, such as VGG and ResNet, in facial expression recognition tasks. The importance of dataset quality, the suitability of different CNN models for various scenarios, and the need for continued contributions to the field are highlighted.

m. An End-to-End Deep Model with Discriminative Facial Features for Facial Expression Recognition:

This paper introduces a hybrid feature representation method that significantly improves recognition accuracy on benchmark datasets. The proposed model's effectiveness encourages future work on face alignment technology to further enhance recognition rates[35].

n. Deep Graph Fusion Based Multimodal Evoked Expressions from Large-Scale Videos:

The deep graph fusion model for video-based emotion prediction exhibits superior performance on the EEV database. Future refinements are suggested to enhance the training approach, considering inter-segment connections, and addressing memory device constraints[36].

o. Deep Learning (DL)-Enabled System for Emotional Big Data:

The emotion care system integrates deep learning with facial expression recognition for autism disorder patient training. Achieving high accuracy and efficient real-time processing, future improvements include model quantification for embedded systems and exploration of real-time applications[37].

p. Dynamic Facial Expression Understanding Using Deep Spatiotemporal LDSP on Spark:

The paper introduces a comprehensive framework for dynamic facial expression recognition using Apache Spark. Future work is planned to explore facial landmark integration and extend the application of the approach to diverse video analysis domains[38].

q. Micro Expression Recognition Using Convolution Patch in Vision Transformer:

The Vision Transformer based on convolution patches effectively addresses challenges in recognizing micro-expressions, showcasing significant improvements in accuracy. Future research directions include addressing unbalanced datasets, exploring in-the-wild datasets, and implementing deep continual learning techniques[39].

r. Multimodal Emotion Recognition From EEG Signals and Facial Expressions

The multimodal model combining EEG and facial expressions achieves high accuracy in emotion recognition. Future research should focus on reliable pre-training models and introduce additional modalities for a more comprehensive understanding of emotions[40].

s. Efficient facial expression recognition algorithm based on hierarchical deep neural network structure:

The hierarchical deep learning approach, fusing appearance and geometric features, showcases improved facial expression recognition accuracy. Future efforts are suggested to enhance recognition speed and explore real-world applications where a hierarchical understanding of facial expressions is beneficial[41].

t. Recognition of Teachers' Facial Expression Intensity Based on Convolutional Neural Network and Attention Mechanism:

The EIDB-13 dataset and the recognition algorithm for teaching quality assessment provide valuable insights into facial expression recognition's applicability beyond traditional emotion analysis. Future work involves improving deep learning robustness and addressing challenges associated with small sample sizes in educational contexts[42].

These discussions collectively highlight the diverse strategies, innovations, and challenges addressed in contemporary facial expression recognition research. The integration of multimodal approaches, lightweight models, and real-world applications showcases the field's ongoing evolution and its potential impact on various domains.

E. Conclusion

The culmination of insights from the reviewed papers underscores the vibrancy and continual evolution within the realm of Facial Expression Recognition (FER). The strides made in this field, fueled by advanced deep learning methodologies, reveal a multifaceted approach to understanding and interpreting facial expressions. From the innovative MBCC-CNN model's effective feature fusion to the Deep Graph Fusion model's utilization of visual and auditory cues, these approaches showcase the multidimensional nature of emotion recognition. The fusion of Electroencephalogram (EEG) signals and facial expressions in the multimodal model signifies a promising avenue for deeper emotional insight. Moreover, the LDSP-TOP descriptor, coupled with Spark-based computations, demonstrates the significance of temporal dynamics in capturing facial expressions. Vision transformers' application in micro-expression recognition adds a layer of sophistication, emphasizing the synergy between local and global information.

The proposed hierarchical deep learning model for teaching quality assessment introduces a new dimension, expanding the scope of FER beyond emotion recognition. Additionally, the visionary transformer model's exceptional accuracy highlights the potential of hybrid architectures, marrying convolutional and transformer networks. In essence, this diverse panorama of FER methodologies not only illuminates the current landscape but also suggests a trajectory of further exploration. As FER continues to play a pivotal role in human-computer interaction, healthcare, and education, the amalgamation of these innovative models points towards a future where nuanced understanding and interpretation of facial expressions become indispensable. The journey from traditional methods to these cutting-edge approaches signifies a paradigm shift, where FER transcends mere recognition and evolves into a comprehensive understanding of human emotions in various contexts.

References

[1] S. M. S. Abdullah and Adnan Mohsin Abdulazeez, "Facial Expression Recognition Based on Deep Learning Convolution Neural Network: A Review," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 53–65, Apr. 2021, doi: 10.30880/jscdm.2021.02.01.006.

[2] G. M. Zebari, D. A. Zebari, D. Q. Zeebaree, H. Haron, Adnan Mohsin Abdulazeez, and K. Yurtkan, "Efficient CNN Approach for Facial Expression Recognition," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Dec. 2021. doi: 10.1088/1742-6596/2129/1/012083.

[3] S. B. Mohammed and Adnan Mohsin Abdulazeez, "Deep Convolution Neural Network for Facial Expression Recognition," 2021.

[4] K. M. Rajesh and M. Naveenkumar, "A robust method for face recognition and face emotion detection system using support vector machines," in *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*, 2016, pp. 1–5. doi: 10.1109/ICEECCOT.2016.7955175.

[5] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020, doi: 10.1109/TIP.2019.2956143.

[6] A. Hassouneh, A. M. Mutawa, and M. Murugappan, "Development of a Real-Time Emotion Recognition System Using Facial Expressions and EEG based on machine learning and deep neural network methods," *Inform Med Unlocked*, vol. 20, p. 100372, Jan. 2020, doi: 10.1016/J.IMU.2020.100372.

[7] Q. Zhu *et al.*, "Learning Temporal and Spatial Correlations Jointly: A Unified Framework for Wind Speed Prediction," *IEEE Trans Sustain Energy*, vol. 11, no. 1, pp. 509–523, 2020, doi: 10.1109/TSTE.2019.2897136.

[8] S. Indolia, S. Nigam, R. Singh, V. K. Singh, and M. K. Singh, "Micro Expression Recognition Using Convolution Patch in Vision Transformer," *IEEE Access*, vol. 11, pp. 100495–100507, 2023, doi: 10.1109/ACCESS.2023.3314797.

[9] Nidhi and B. Verma, "From methods to datasets: a detailed study on facial emotion recognition," *Applied Intelligence*, vol. 53, no. 24, pp. 30219–30249, 2023, doi: 10.1007/s10489-023-05052-y.

[10] E. K. Babu, K. Mistry, M. N. Anwar, and L. Zhang, "Facial Feature Extraction Using a Symmetric Inline Matrix-LBP Variant for Emotion Recognition," *Sensors*, vol. 22, no. 22, Nov. 2022, doi: 10.3390/s22228635.

[11] S. Li, M. Yuan, J. Chen, and Z. Hu, "AdaDC: Adaptive Deep Clustering for Unsupervised Domain Adaptation in Person Re-Identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3825–3838, 2022, doi: 10.1109/TCSVT.2021.3118060.

[12] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, vol. 91, pp. 424–444, Mar. 2023, doi: 10.1016/J.INFFUS.2022.09.025.

[13] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal Intelligence: Representation Learning, Information Fusion, and Applications," *IEEE J Sel Top Signal Process*, vol. 14, no. 3, pp. 478–493, 2020, doi: 10.1109/JSTSP.2020.2987728.

[14] A. Jalal, A. Ahmed, A. A. Rafique, and K. Kim, "Scene Semantic Recognition Based on Modified Fuzzy C-Mean and Maximum Entropy Using Object-to-Object Relations," *IEEE Access*, vol. 9, pp. 27758–27772, 2021, doi: 10.1109/ACCESS.2021.3058986.

[15] M. A. Uddin, J. B. Joolee, and K.-A. Sohn, "Dynamic Facial Expression Understanding Using Deep Spatiotemporal LDSP On Spark," *IEEE Access*, vol. 9, pp. 16866–16877, 2021, doi: 10.1109/ACCESS.2021.3053276.

[16] L. Zhang, X. Hong, O. Arandjelović, and G. Zhao, "Short and Long Range Relation Based Spatio-Temporal Transformer for Micro-Expression Recognition," *IEEE Trans Affect Comput*, vol. 13, no. 4, pp. 1973–1985, 2022, doi: 10.1109/TAFFC.2022.3213509.

[17] H. Zhou, S. Huang, and Y. Xu, "Inceptr: micro-expression recognition integrating inception-CBAM and vision transformer," *Multimed Syst*, vol. 29, no. 6, pp. 3863–3876, 2023, doi: 10.1007/s00530-023-01164-0.

[18] H. Sadeghi, A.-A. Raie, and M.-R. Mohammadi, "Facial expression recognition using geometric normalization and appearance representation," in *2013 8th Iranian Conference on Machine Vision and Image Processing (MVIP)*, 2013, pp. 159–163. doi: 10.1109/IranianMVIP.2013.6779970.

[19] J. Lee, S. Kim, S. Kim, and K. Sohn, "Multi-Modal Recurrent Attention Networks for Facial Expression Recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 6977–6991, 2020, doi: 10.1109/TIP.2020.2996086.

[20] J.-H. Kim, B.-G. Kim, P. P. Roy, and D.-M. Jeong, "Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure," *IEEE Access*, vol. 7, pp. 41273–41285, 2019, doi: 10.1109/ACCESS.2019.2907327.

[21] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, "Learning Personalized Models for Facial Expression Analysis and Gesture Recognition," *IEEE Trans Multimedia*, vol. 18, no. 4, pp. 775–788, 2016, doi: 10.1109/TMM.2016.2523421.

[22] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning From Hierarchical Spatiotemporal Descriptors for Micro-Expression Recognition," *IEEE Trans Multimedia*, vol. 20, no. 11, pp. 3160–3172, 2018, doi: 10.1109/TMM.2018.2820321.

[23] D. Shehada, A. Turky, W. Khan, B. Khan, and A. Hussain, "A Lightweight Facial Emotion Recognition System Using Partial Transfer Learning for Visually Impaired People," *IEEE Access*, vol. 11, pp. 36961–36969, 2023, doi: 10.1109/ACCESS.2023.3264268.

[24] D. H. Lee and J. H. Yoo, "CNN Learning Strategy for Recognizing Facial Expressions," *IEEE Access*, vol. 11, pp. 70865–70872, 2023, doi: 10.1109/ACCESS.2023.3294099.

[25] J. Pu and X. Nie, "Convolutional Channel Attentional Facial Expression Recognition Network and Its Application in Human-Computer Interaction," *IEEE Access*, vol. 11, pp. 129412–129424, 2023, doi: 10.1109/ACCESS.2023.3333381.

[26] Y. Zhou, Y. Liang, and P. Tan, "Design of an Intelligent Laboratory Facial Recognition System Based on Expression Keypoint Extraction," *IEEE Access*, vol. 11, pp. 129805–129817, 2023, doi: 10.1109/ACCESS.2023.3329575.

[27] M. Hussain, E. U. H. Qazi, H. A. Aboalsamh, and I. Ullah, "Emotion Recognition System Based on Two-Level Ensemble of Deep-Convolutional Neural Network Models," *IEEE Access*, vol. 11, pp. 16875–16895, 2023, doi: 10.1109/ACCESS.2023.3245830.

[28] J. Lan *et al.*, “Expression Recognition Based on Multi-Regional Coordinate Attention Residuals,” *IEEE Access*, vol. 11, pp. 63863–63873, 2023, doi: 10.1109/ACCESS.2023.3285781.

[29] Y. He, “Facial Expression Recognition Using Multi-Branch Attention Convolutional Neural Network,” *IEEE Access*, vol. 11, pp. 1244–1253, 2023, doi: 10.1109/ACCESS.2022.3233362.

[30] H. M. Shahzad, S. M. Bhatti, A. Jaffar, M. Rashid, and S. Akram, “Multi-Modal CNN Features Fusion for Emotion Recognition: A Modified Xception Model,” *IEEE Access*, vol. 11, pp. 94281–94289, 2023, doi: 10.1109/ACCESS.2023.3310428.

[31] G. Perveen *et al.*, “Multi-Stream Deep Convolution Neural Network With Ensemble Learning for Facial Micro-Expression Recognition,” *IEEE Access*, vol. 11, pp. 118474–118489, 2023, doi: 10.1109/ACCESS.2023.3325108.

[32] B. Fang, X. Li, G. Han, and J. He, “Rethinking Pseudo-Labeling for Semi-Supervised Facial Expression Recognition With Contrastive Self-Supervised Learning,” *IEEE Access*, vol. 11, pp. 45547–45558, 2023, doi: 10.1109/ACCESS.2023.3274193.

[33] C. Shi, C. Tan, and L. Wang, “A Facial Expression Recognition Method Based on a Multibranch Cross-Connection Convolutional Neural Network,” *IEEE Access*, vol. 9, pp. 39255–39274, 2021, doi: 10.1109/ACCESS.2021.3063493.

[34] L. V. L. Pinto *et al.*, “A Systematic Review of Facial Expression Detection Methods,” *IEEE Access*, vol. 11, pp. 61881–61891, 2023, doi: 10.1109/ACCESS.2023.3287090.

[35] J. Liu, H. Wang, and Y. Feng, “An End-to-End Deep Model with Discriminative Facial Features for Facial Expression Recognition,” *IEEE Access*, vol. 9, pp. 12158–12166, 2021, doi: 10.1109/ACCESS.2021.3051403.

[36] N. H. Ho, H. J. Yang, S. H. Kim, G. Lee, and S. B. Yoo, “Deep Graph Fusion Based Multimodal Evoked Expressions from Large-Scale Videos,” *IEEE Access*, vol. 9, pp. 127068–127080, 2021, doi: 10.1109/ACCESS.2021.3107548.

[37] H. Wang, D. P. V. Tobón, M. S. Hossain, and A. El Saddik, “Deep Learning (DL)-Enabled System for Emotional Big Data,” *IEEE Access*, vol. 9, pp. 116073–116082, 2021, doi: 10.1109/ACCESS.2021.3103501.

[38] M. A. Uddin, J. B. Joolee, and K. A. Sohn, “Dynamic Facial Expression Understanding Using Deep Spatiotemporal LDSP on Spark,” *IEEE Access*, vol. 9, pp. 16866–16877, 2021, doi: 10.1109/ACCESS.2021.3053276.

[39] S. Indolia, S. Nigam, R. Singh, V. K. Singh, and M. K. Singh, “Micro Expression Recognition Using Convolution Patch in Vision Transformer,” *IEEE Access*, vol. 11, pp. 100495–100507, 2023, doi: 10.1109/ACCESS.2023.3314797.

[40] S. Wang, J. Qu, Y. Zhang, and Y. Zhang, “Multimodal Emotion Recognition From EEG Signals and Facial Expressions,” *IEEE Access*, vol. 11, pp. 33061–33068, 2023, doi: 10.1109/ACCESS.2023.3263670.

[41] J. H. Kim, B. G. Kim, P. P. Roy, and D. M. Jeong, “Efficient facial expression recognition algorithm based on hierarchical deep neural network structure,” *IEEE Access*, vol. 7, pp. 41273–41285, 2019, doi: 10.1109/ACCESS.2019.2907327.

[42] K. Zheng, D. Yang, J. Liu, and J. Cui, “Recognition of teachers&#x2019; facial expression intensity based on convolutional neural network and attention mechanism,” *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.3046225.