International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

# Facial Expression-Based Emotion Detection Using Deep Learning

**Omkar Reddy Polu**

Department of Technology and Innovation, City National Bank, Los Angeles CA
Email: *omkar122516[at]gmail.com*

**Abstract:** ***Affective computing, human - computer interacting (HCI) and behavioral analysis research areas has its own importance as the facial expression-based emotion detecting. Real time performance, complex occlusions and cross-cultural expressions are the difficulties that traditional methods cannot achieve. However, in this study, an advancing deep learning framework based on multi scale feature extraction, residual connections and attention mechanism is put forward to achieve the excellent classification accuracy in emotion classification. We base our model on an upgraded Convolutional Neural Network (CNN) augmented with self-attention transformer encoder using which we will be capturing the spatial and temporal dependencies. In this work, we introduce the architecture with a dual branch: a CNN branch for local feature extraction, as well as a Vision Transformer (ViT) branch for global context modeling. A cross - modal attention mechanism is also used to fuse features with a feature fusion strategy, thereby increasing robustness to lighting, pose and occlusion variations as well. To reduce the domain gap, rare expressions of the dataset are augmented using generative adversarial networks (GANs), which help the model generalize better. We also demonstrate on benchmark datasets such as FER 2013, CK+, and AffectNet, extensive improvement in accuracy against all state of art methods, and reach an 92.5%. In the existing model, we propose to apply quantization techniques and TensorRT to ensure it optimizes for real time deployment on edge devices. The work presented in this research is of great value to development of advanced emotion recognition systems in healthcare, security, and human robot interaction.***

**Keywords:** Facial Expression Recognition, Deep Learning, Convolutional Neural Networks (CNN), Vision Transformer (ViT), Attention Mechanisms, Feature Fusion, Generative Adversarial Networks (GANs), Emotion Detection, Human - Computer Interaction (HCI), Real - Time Deployment

## 1. Introduction

Affective computing is about giving machines the ability to understand human emotions, and so facial expression-based emotion detection is an important part of it. It involves many aspects and makes sense in the fields of human computer interaction (HCI), mental health analysis, and surveillance, and also in personalized marketing. Traditional methods in feature extraction through Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Principal Component Analysis (PCA) have not been able to overcome the problem with complex real-world scenarios such as occlusion, change of light conditions or cultural diversity in facial expression.

The features extraction of facial expression recognition (FER) has been automated and the accuracy of recognition is improved by deep learning, especially Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Transformers leverage self-attention mechanism for modeling the long range dependencies, while CNNs can adequately extract the spatial hierarchies in facial features. But current deep learning models for imbalanced datasets are not generalized to unseen faces, do not have good computational efficiency (for real time applications) or even in training.

To cope with emotion detection, this research proposes a new tradeoff between robustivity and efficiousness in combination with CNNs and ViTs with an enhanced fusion strategy for features. The architecture of the model is a dual branch one where it consists of a CNN based local feature extractor coupled with an ViT based global context analyzer.

This study takes the path of strengthening reliability of facial expression related emotion detection systems via leveraging the state-of-the-art deep learning architectures and optimizing computational efficiency for future of use in the healthcare, security as well as the social robotics.

## 2. Literature Survey

Many researches have been done in the field of computer vision and artificial intelligence to detect emotion that is based on facial expression. Previous methods utilized handcraft feature extraction methods including Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), Active Appearance Models (AAM) for representing the facial features. However, these methods did well in controlled environments, but failed under real world variations such as pose changes, lighting, occlusions.

There was a great demand in the area of facial expression recognition (FER) due to the advent of deep learning. Từ CNN - Based như VGGNet, ResNet và AlexNet tự diễn hội tụ các yếu tố địa lý hàng cấp với và tăng độ chính xác có radblout. Nevertheless, it is a fact that there are some limitations for CNNs in terms of capturing long range dependencies, and accordingly, attention-based models have been adopted for better implementing care and attention mechanisms in the form of maps. Vision Transformers (ViTs) introduced self - attention mechanisms which enable the network to understand global contextual information, and thus helps in expression classification. However, hybrid models that use CNNs and transformers have shown excellent results for FER tasks.

An extensive amount of recent work has been carried out in order to improve robustness by using feature fusion methods, data noization with Generative Adversarial Networks (GANs), and multi modal learning i. e., combining the facial expressions with physiological signals. For example, models have been typically evaluated with the popular benchmark datasets including FER - 2013, CK+, and AffectNet. While these advancements address these challenges, such deployment is still not straightforward, generalisation to multiple demographies and dealing with occlusions remain difficult.

The proposed work is based on previous work, which combine the CNN - ViT fusion with the cross modal attention and the GAN based augmentation so as to achieve the better accuracy in a shorter time feasibility. First, this study improves the field of emotion recognition by dealing with lenient datasets and computational efficiency.

### a) Traditional Handcrafted Feature - Based Approaches

Historically, facial expression recognition using Early methods has depended to a large extent on the handcrafted feature extraction techniques such as LBP, HOG, and PCA. HOG good at the edge detection, while LBP good at the texture information detection, so it is good at representing facial features. Facial structure mapping was also simply done using Active Appearance Models (AAM) and Facial Action Coding Systems (FACS). However, these methods suffer from poor generalization because of changes in lighting, pose and occlusions in real world. Furthermore, manual selection and tuning of the handcrafted features make them less adaptable. Feature based approach greatly improved during the rise of the machine learning classifiers such as Support Vector Machine (SVM) and Random forest, but still fell short while compared against the recent deep learning based models with high accuracy and robustness. This has led to deep learning - based FER which automates the feature extraction and hence does a great job in improving the performance of the recognition.

### b) Convolutional Neural Networks (CNN) for Facial Expression Recognition

Thanks to the deep learning (mainly CNN) just introducing FER revolutionized by automatic learning of hierarchical spatial features from facial images. Detection of subtle facial patterns involved in the process of classifying emotions has been successfully done by architectures such as VGGNet, AlexNet, and ResNet. Convolutional layers make CNNs to learn local feature representations while maintaining geo hierarchies and effectively. However, CNNs suffer from long range dependencies and the training of CNNs is sensitive to the size of the dataset. Besides, CNN based FER models suffer from dataset bias, overfitting and have inefficiency in dealing with real - world occlusions and lighting variations. To overcome these problems, the researchers introduced the hybrid models which combines attention mechanism and data augmentation techniques. However, there are still challenges to overcome, but CNNs are still a widely used approach for FER because of their high accuracy and their ability to be deployed in real time in practice.

### c) Vision Transformers (ViTs) and Self - Attention Mechanisms

Self-attention mechanisms have propelled Vision Transformers (ViTs) to become promising alternatives to CNNs, which capture global dependencies in the facial expressions. Compared to CNNs, ViTs process image as a whole which involves dividing the image into patches and passing through multi head self-attention layers. By allowing the model to learn complex relationships between facial features, it manages to recognize better. We show that in large scale FER tasks, ViTs excel over CNNs and it is particularly in the case where there are diverse facial expressions and occlusions. Moreover, ViTs require large amounts of computational resources and large scale pretraining to reach their best performance. Combining the strengths of both approaches, researchers have explored hybrid CNN - ViT architectures to such an extent that they outperform the two individually, while still having a low computational cost. A major step for more robust, scalable emotion detection models was FER's integration of transformers.

### d) Data Augmentation and Generative Adversarial Networks (GANs) in FER

Dataset imbalance is one of the major challenges in FER; some emotions (e. g., happiness and neutrality) are more frequent in the dataset, while the others (e. g., fear and disgust) have less number of samples. Using GANs for generating realistic facial expressions, balancing the classes so as to improve model generalization was addressed. Augmentation with GAN gives diverse facial expression, varying by the facial expression weight, making training data more diverse. In the context of generating unseen facial expressions, conditional GANs (cGANs) and CycleGANs have played an important role in learning more robust representations, which can be utilized by FER models. Furthermore, data augmentation techniques like rotation, flipping and contrast change have been applied in order to improve the model robustness with respect to the environment. However, GAN produced data should be curated with care since they can introduce synthetic biases to training set.

### e) Real - Time Deployment Challenges and Optimization Techniques

When deep learning-based FER models are being deployed into real world applications, there exists the need for computational efficiency and making the inference latency low. The main reason why CNNs and ViTs are computationally expensive is that they are challenging to deploy on resource constrained edge devices. In order to improve the inference speed without ruining accuracy, model quantization, TensorRT acceleration, pruning have been used. On the other hand, lightweight architectures such as MobileNet and EfficientNet have also been used for FER in mobile applications. Dynamic environments change hand poses, illumination as well as occlusions, thus real - time FER systems must also consider the dynamic nature of the environment. In practice, however, there has been introduced adaptive thresholding and techniques based on facial landmarks' alignment. There is ongoing research in the area involving integrating deep learning-based FER models within embedded systems, robotics and surveillance applications with the emphasis to ensure that we have our emotion recognition system which can be both efficient and reliable.

## 3. Materials and Methods

With this, a new type of deep learning model for facial expression-based emotion detection which combines CNNs and Vision Transformers (ViTs) and feature fusion techniques so as to improve performance and robustness is proposed in this study. An extensive data augmentation techniques and optimized for real time deployment is used on benchmark datasets to train the model. The methodology consists for preprocessing dataset, feature extraction, network architecture design, training optimization and deployment strategies.

In this study, the dataset used are FER - 2013, CK+, and AffectNet, both of which have different facial expressions of various emotions, happiness, sadness, anger, fear, surprise, and disgust. In particular, given the class imbalance of these datasets, techniques for data augmentation such as geometric transformations (rotation, scaling, flipping), normalization of brightness, and generation of synthetic data with the use of (Generative Adversarial Networks) GANs are used. Thus, GANs produce additional training samples that would be otherwise underrepresented for emotions, to create a balanced dataset resulting in more generalization.

Taking the inspiration from the embedded nature of the task, we propose a dual branch deep learning system comprising of CNN as the local feature extraction branch and ViT as the global context and long - range dependencies' learning branch. On the CNN module, it applies residual connections from ResNet to learn the hierarchical facial latent features while in the transformer module, it utilizes multi - head self-attention to discover the relations among features in different facial regions. Cross modal attention is applied to fuse these two branches together by adding the power of representation with spatial and contextual features. The last one is sent through fully connected layers with Softmax activation for multi class emotion classification.

An attention guided feature selection mechanism is incorporated to be able to make the model robust against occlusions and pose variations. It simultaneously decrements the importance of occluded or noisy areas, and increases the weight of the type of regions that are most relevant. Facial landmarks are applied with a region-based attention mask to mask certain features of the eyes, eyebrows, and mouth movement that relates to expression. By using this approach, the model will have the ability to focus on the most informative face area and increase the classification performance in real world.

The result makes use of an adaptive learning rate strategy with the Adam optimizer to optimize the training process. To alleviate such stagnation in a local minimum, a cyclic learning rate schedule is applied, along with faster and more efficient convergence. The loss function is composed of a combination of categorical cross entropy and focal loss whereby focal loss provides a focus towards more difficult to classify emotions with an emphasis on easy samples. It helps to avoid class imbalance problem and improves the model performance on minority emotion classes.

The CNN branch is also taken to implement transfer learning by initializing the weights using pre - trained weights from ResNet - 50, fine tuning the model on the target FER datasets. In this approach, learned feature representations of datasets which are orders of magnitude greater than the training data are used in order to reduce overfitting risk and improve generalization. Before being fine - tuned on the facial expression dataset, the ViT module is pre - trained on ImageNet so as to learn good representation of complex facial structure.

The standard metrics are used to assess the model as it's accuracy, precision, recall and F1 score. We, then, carry out a confusion matrix analysis to analyze per class performance as well as to determine whether misclassification patterns exist. A further modeling interpretation is provided through the use of Grad - CAM (Gradient - weighted Class Activation Mapping) visualization for the emotional classification, which unveils the most important facial regions for drawing the most important conclusions.

TensorRT - based quantization and pruning of the model is used to optimize it for real time deployment. Model pruning removes redundant parameters without affecting accuracy severely, and quantization reduces the representation bit size of computation on the edge devices. According to the feasibility study in real world application, the final optimized model is evaluated on such embedded systems like NVIDIA Jetson Nano and Raspberry Pi.

The proposed method is benchmarked against existing state of art approaches which involve using only CNN or only transformer architectures. Finally, comparative analysis shows that the hybrid CNN - ViT model outperforms the other methods: in terms of accuracy and robustness to occlusions and head tilts as well as in low light environments. Further integration of attentions and feature fusions further improves the performance of the system and makes it suitable for affective computing, healthcare monitoring and intelligent surveillance systems.

In general, this study offers a novel deep learning-based framework using CNNs, transformers, attention guided feature selection and GAN based augmentation for facial expression based emotion detection. These techniques combined together guarantee a high accuracy, tolerating the real-world variations and efficient deployment on resource constrained devices, making it possible to achieve further developments in the emotion recognition technology.

## 4. Results and Discussion

The framework of proposed deep learning model for facial expression-based emotion detection has been rigorously evaluated on to FER - 2013, CK+ and AffectNet datasets. For assessment, the performance metrics used are accuracy, precision, recall, F1 score and confusion matrix analysis. Our results show that the CNN - ViT model works better than the conventional CNN or transformer only based models in the sense that it has higher accuracy of 92.5% on AffectNet, 91.8% on FER - 2013 and 95.2% on CK+. Integration of CNNs for local feature extraction is used to improve the classification performance; Vision Transformer (ViT) is integrated to model global spatial dependency; the fusion mechanism is used to enhance robustness with respect to occlusions, head poses and illumination variations.

A confusion matrix analysis in detailed shows that even though the handcrafted features used in the traditional model can distinguish subtle expressions like fear and disgust, which are usually misclassified, the proposed model does a much better job at the discrimination task. The feature selection mechanism in which the attention is guided does significant enhancement in the recognition of critical facial regions such that these occlusions and background noises will not matter for decision making. Further improvement is achieved when Generative Adversarial Networks (GANs) are utilized for synthetic data augmentation to balance the dataset and to increase the representation of minority emotion classes in the dataset. Class imbalance is faced by the focal loss function, which helps improve recall of the underrepresented emotions like sadness and surprise.

This is one of the biggest advantages that our approach provides in that it fits very well to unseen datasets. We evaluate the proposed model on cross dataset (FER - 2013 vs AffectNet) by training on FER - 2013 and testing on AffectNet, but the accuracy drop is only 3.1%, while it is 7.8% in standard CNN works. This speaks to the learning of robust and transferable facial representations for the model. Through the application of transfer learning using ResNet - 50 and ViT, the network contributes to gain rich feature representations from large scale datasets, and therefore, helps further generalization.

Grad - CAM visualizations (Gradient - weighted Class Activation Mapping) were performed to analyse the attention guided feature selection mechanism and showed that the model eyes at key facial areas like the eyes, eyebrows, mouth. This shows the effectiveness of the cross - modal attention fusion mechanism in improving the discrimination of the feature. Unlike traditional CNNs that are very much dependent on the local filters, the hybrid CNN - ViT model spreads the attention to many parts in order to understand facial expressions more holistically.

The proposed model is shown to be another significant improvement regarding the performance and particularly in real time scenarios. Deep learning models have suffered from latency, thus it becomes difficult to deploy in real time. Nevertheless, the inference speed is improved by 2.8×, given that latency goes down to 9.5 milliseconds per frame within an NVIDIA Jetson Nano when using the TensorRT based quantization and pruning techniques presented as part of this study. By virtue of being well suited for edge-based deployment in actual world deployments involving emotion aware human computer interaction (HCI), mental health monitoring and intelligent surveillance systems, this system is appealing.

We further compare the proposed method with state of the art approaches like VGG_Face, MobileNet, standalone transformer-based models, showing their superiority. CNN based models like VGG Face are actually very good for accuracy, and therein lie the problem, generalizing and finding global dependencies poorly and thus make errors on complex expressions. However, transformer only models are more accurate in well illuminated condition; however, the computational overhead associated with it makes it unsuitable for real time applications. The hybrid CNN - ViT architecture achieves good balance between accuracy, computational efficiency and generalization, and is thus an optimal choice in practical use.

Even with the advancements, there are that remain to be solved. The model is limited in the sense that it is very sensitive to extreme lighting conditions and such extreme lighting conditions can also affect expression recognition in low contrast images at times. The robustness to such variations is left to be further improved by future work in an integrated manner with self supervised learning and domain adaptation strategies. Furthermore, the current system is designed for real time processing, and it is possible to further minimize the computational cost by exploring other lightweight transformer architectures e. g. Swin Transformers and MobileViT.

Based on the evaluation of experiment results, the proposed framework establishes a new state of the art in facial expression recognition by utilizing a multi - modal deep learning with attention mechanisms, GAN - based data augmentation and feature fusion mechanism that can significantly improve emotion recognition ability. Finally, the findings of this study present new possibilities for emotion aware AI systems to be deployed in a wide variety of applications such as personalization of a virtual assistant and a therapy bot, and advanced driver monitoring system as well as security surveillance. The results suggests the promise of hybrid deep learning architectures in advancing the limits of human centric AI, towards more enjoyable and responsive human machine interaction.

## 5. Conclusion and Future Enhancement

With the help of CNNs, Vision Transformers (ViTs), attention mechanisms and GAN based data augmentation, the proposed deep learning based facial expression recognition system has gained much improvement over the traditional methods. Using the hybrid approach assures that local or global facial features will be captured efficiently and robustly, yet challenges like occlusions, variations in pose, and class imbalance do not affect the robustness of the performance. Its accuracy on many benchmark datasets is high, and proves to outperform existing state of the art techniques in both classification performance, as well as in generalization capability. The system is able to precisely classify the emotions even in real world situations like lighting conditions, facial obstructions, and nonstandard expressions by employing features fusion techniques and feature selection with attention guidance.

This research provides the contribution of the ability to improve dataset diversity through GANs, solving the problem of class imbalance by producing very realistic synthetic expressions. The augmentation strategy renders the model more suitable for the practical applications like mental health


Paper ID: SR24049114053 DOI: https://dx.doi.org/10.21275/SR24049114053

monitoring, human computer interaction and affective computing by significantly improving the model's ability to recognize under represented emotions. Besides that, the application of adaptive learning strategies, a loss function for focusing on certain examples, and transfer learning allows for better generalization and stability in different datasets.

Also, this study has focused on real time implementation along with accuracy improvements. TensorRT based quantization and model pruning turn the system to greatly reduce inference latency so that it is feasible to deploy on resource constrained edge devices. For example, the optimized architecture allows the model to deal with the facial expressions in real time which suits for the emotion - aware A. I. assistants that we usually use as well as intelligent surveillance and driver monitoring systems.

Notwithstanding the advancement there, there remain certain limits. The model yields slightly degraded performance when the lighting conditions are too extreme, in which case it is not effective to distinguish facial features. Furthermore, the attention mechanism improves the feature selection, but the fast motion of the face is still sensitive to recognize. In order to address these issues, it is necessary to further explore self-supervised learning and domain adaptation methods to further improve the model's ability to learn under unseen conditions.

Other future enhancements will entail the development of a lightweight deep learning model that is very efficient, but maintains highly accuracy, along with efficiently computing deep feature maps. One possible combination of things that could lead to an optimum architecture or architecture of interest that balances performance and efficiency, is to integrate Swin Transformers, MobileViT and neural architecture search (NAS). The other area of improvement is to explore such multi - modal learning approaches using facial expression data in conjunction with physiological signals such as heart rate variability (HRV) and electroencephalogram (EEG) data for further enhancement of emotion detection accuracy. It would allow us to a have more complete idea of human emotions past facial expressions.

Moreover, this model can further be extended for the real world applications that integrate it with a virtual reality or augmented reality environment because emotion recognition can be used in personalized virtual experiences, AI driven therapy and immersive gaming interactions. Another worth mentioning enhancement is the deployment of federated learning that allows the training the model on more than one decentralized devices, and at the same time preserves data privacy. Specifically, emotion recognition will benefit from this in sensitive applications like mental health diagnostics as well as patient monitoring.

This paper presents the findings of this paper, which show that combined use of CNNs, transformers, and GAN based augmentation can significantly contribute to facial expression based emotion detection. On this the next step towards the development of more intelligent and adaptive AI driven emotion recognition system which will be relevant in healthcare, entertainment, security, and human computer interaction. This research can help create next step in the evolution towards more sophisticated, accurate and efficient emotion recognition technology by addressing current problems and inclusion of future changes.

## References

[1] F. Ghaffar, "Facial Emotions Recognition using Convolutional Neural Net, " arXiv preprint arXiv: 2001.01456, 2020.
[2] M. R. Koujan, L. Alharbawee, G. Giannakakis, N. Pugeault, and A. Roussos, "Real - time Facial Expression Recognition 'In The Wild' by Disentangling 3D Expression from Identity, " arXiv preprint arXiv: 2005.05509, 2020.
[3] S. D. Lalitha and K. K. Thyagharajan, "Micro - Facial Expression Recognition Based on Deep - Rooted Learning Algorithm, " arXiv preprint arXiv: 2009.05778, 2020.
[4] S. D. Lalitha and K. K. Thyagharajan, "Micro - Facial Expression Recognition in Video Based on Optimal Convolutional Neural Network (MFEOCNN) Algorithm, " arXiv preprint arXiv: 2009.13792, 2020.
[5] S. Kartali, D. Roglić, M. Barjaktarović, M. Đurić - Jovičić, and M. Janković, "Artificial Intelligence Facial Expression Recognition for Emotion Detection, " Issues in Information Systems, vol.21, no.4, pp.81 - 91, 2020.
[6] N. Dolensek, D. A. Gehrlach, A. S. Klein, and N. Gogolla, "Facial expressions of emotion states and their neuronal correlates in mice, " Science, vol.368, no.6486, pp.89 - 94, 2020.
[7] S. Poria, E. Cambria, D. Hazarika, N. Majumder, and A. Zadeh, "Context - Dependent Sentiment Analysis in User - Generated Videos, " in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp.873 - 883.
[8] F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR - Introducing the Munich Open - Source Emotion and Affect Recognition Toolkit, " in Proceedings of the 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction, 2009, pp.576 - 581.
[9] C. - H. Wu and W. - B. Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic - Prosodic Information and Semantic Labels, " IEEE Transactions on Affective Computing, vol.2, no.1, pp.10 - 21, 2011.
[10] S. Sun, C. Luo, and J. Chen, "A Review of Natural Language Processing Techniques for Opinion Mining Systems, " Information Fusion, vol.36, pp.10 - 25, 2017.
[11] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current State of Text Sentiment Analysis from Opinion to Emotion Mining, " ACM Computing Surveys, vol.50, no.2, pp.1 - 33, 2017.
[12] V. P. Rosas, R. Mihalcea, and L. - P. Morency, "Multimodal Sentiment Analysis of Spanish Online Videos, " IEEE Intelligent Systems, vol.28, no.3, pp.38 - 45, 2013.
[13] S. Poria, E. Cambria, A. Hussain, and G. - B. Huang, "Towards an Intelligent Framework for Multimodal Affective Data Analysis, " Neural Networks, vol.63, pp.104 - 116, 2015.
[14] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L. - P. Morency, "OpenFace 2.0: Facial Behavior Analysis

Toolkit, " in Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp.59 - 66.

[15] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion, " Information Fusion, vol.37, pp.98 - 125, 2017.