Design and Implementation of a Compact Convolution Neural Network for Real Time Facial Emotional Recognition

Sahar Afzal, Deepti Razdan, Bharat Singh Ladhi
Dept of Computer Application
Eklavya University
Damoh,M.P
afzal.sahar@yahoo.com ,razdandeepti09@gmail.com , bslodhi02@gmail.com

How to Cite this Article:

Afzal, S., Razdan, D. & Ladhi, B. S. (2026). Design and Implementation of a Compact Convolution Neural Network for Real Time Facial Emotional Recognition. International Journal of Creative and Open Research in Engineering and Management, <i>02</i>(8), 1-9. https://doi.org/10.55041/ijcope.v2i8.151

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.

https://doi.org/10.55041/ijcope.v2i8.151

Abstract—Facial expressions are crucial signals that convey messages, exchange thoughts, and express emotions through subtle facial movements. Recognizing human emotions remains a difficult endeavour since people frequently exhibit the same internal states in different ways, affected by cultural differences and personal habits. Modern deep learning architecture particularly the Enhanced Compact Convolution Transformer provide remarkable superior result for detecting a facial emotions and give best result as compare to convolution neural network. This research recognise facial expression by using deep learning architecture utilize both local and global features. They are designed as hybrid models that bridge the gap between the local feature extraction of CNNs and the global context modelling of vision transformers. In this research we identify seven different stages of emotions like angry,disgust,Fear,sad,Happy, surprise,Natural.This experiment is conducted in FER2013 dataset. The proposed model achieve an highest accuracy of as compare to Vision Transformer,Compect Convolutional Transformer

Keywords- *Convolution Neural Network, Multi-Scale Convolution Tokenization, Adaptive Attention Module, Compact Convolution Transformer.*

I. INTRODUCTION

In contemporary technology, human-machine contact is becoming more and more crucial. Machines need to be able to comprehend human emotions and actions in order to communicate with people in an efficient manner. A machine that can identify human emotions will be better able to comprehend human behavior and give users insightful feedback that will increase efficiency and productivity. Emotion shapes almost every part of our daily life ,driving how you think and communicate, Only 7% of meaning in face-to-face communication comes from spoken words, 38% from tone of voice, and 55% from facial expressions, according to a 1968 study by Albert Mehrabian[1].This result emphasizes how crucial facial expressions are to comprehending human emotions. Therefore, facial expression analysis has become one of the most essential methodologies for emotion recognition[2].Developing trustworthy real-time emotion identification systems is still difficult, despite extensive research on face expression recognition from 2D photos. Poor image quality, non-frontal face positions, and shifting lighting conditions are common in real-world settings. These differences make it more challenging to recognize emotions, and additional study is needed to create systems that can reliably identify emotions in such circumstances. [3].The process of recognizing human emotions is known as emotion recognition. Individuals differ in their capacity to accurately identify others' emotions. Due to developments in deep learning and artificial intelligence (AI), which enable machines to recognize emotions automatically, this topic has received a lot of interest lately. The creation of automated systems that can identify emotions has long piqued the curiosity of researchers [2].Nowadays, a variety of information sources can be used to recognize emotions. These include reviewing data from social media platforms, analyzing face expressions in photos and videos, and researching voice signals in audio recordings. Additionally, with the aid of AI algorithms, physiological markers including brain activity, ECG signals, and body temperature are increasingly being employed for emotion identification. Deep learning techniques can assist businesses in marketing and advertising by analyzing consumers' emotional responses to ads, products, and designs. Businesses can improve their marketing tactics and boost sales by using this information to target adverts to those who are more likely to be interested in particular products [3].Emotion recognition systems are used in education to track students' interest in and participation in class activities. Based on students' emotional reactions, these systems can also offer input for the creation of individualized learning resources. Emotion detection can be used in healthcare and wellness to track stress levels and suggest mindfulness or relaxation techniques. Security systems can also benefit from emotion recognition [4]. Before dangerous acts take place, violent or suspicious behavior can be recognized with the aid of real-time emotion detection. In general, emotion detection systems have many uses in fields including marketing, education, healthcare, and security and are crucial to enhancing human-machine interaction[5].

II. RESEARCH OBJECTIVES

The main goal of this research is to use hybrid deep learning techniques to create an accurate and efficient emotion recognition system. The following points are the objective of this research

(i)Develop a real time emotional recognition with hybrid model.

(ii)Design a compact transformer model for comparison with convolution models.

(iii)Experiment is conducted in FER2013 dataset to assess both models performance.

(iv)To evaluate how well hybrid systems capture both local and global features.

III. RELATED WORK

Ozdemir et al. [6] introduced a CNN-based LeNet framework for recognizing emotions from facial expressions, applicable in fields such as biomedical engineering, psychology, and neurology. By combining three datasets (JAFFE, KDEF, and a proprietary dataset), they trained the model to recognize seven types of emotions from facial expressions. This research has the potential to assess mental health conditions and understand customer behavior by analyzing facial expressions in real-time. The CNN structure, optimized for enhanced efficiency, demonstrates its utility across various domains and provides valuable insights for future research.

Xiang and others [7] have discussed in detail the new use of Multi-Task Cascaded Convolutional Networks (MTCNN) for detecting faces and identifying facial features together. They emphasize the coordination between these tasks, which is often overlooked, and explain how much potential there is for improvement. This article highlights facial expression recognition (FER) and its development significance, mentioning key events like Ekman and others' 1971 study on six common facial expressions and the use of Convolutional Neural Networks (CNNs). Adopting the MTCNN originally created by Kaipeng Zhang and others, the authors have developed a unique approach for both face detection and FER. Their tests using the FER2013 dataset achieved 60.7% validation accuracy, and they plan to improve accuracy and performance further in future versions by adding more layers and filters.

Nwosu and others [8] introduced a new method in their paper, "Deep Convolutional Neural Network for Facial Expression Recognition using Facial Parts." Their approach aims to improve facial expression recognition (FER) systems using deep convolutional neural networks (CNNs) by taking facial parts as input, which enhances both performance and speed. The process involves image pre-processing, extracting facial features using Viola-Jones face detection, and using the eye and mouth regions for CNN-based classification.

Transformer-based architectures which shown in fig 1 have received a lot of interest lately in computer vision and natural language processing. Large-scale pre-training followed by task-specific fine-tuning allowed models like BERT and GPT to reach state-of-the-art performance in natural language processing. However, the quadratic complexity of the self-attention mechanism for processing high-resolution images presents major computational hurdles when transformers are used directly to visual data [9].

Convolutional Neural Networks (CNNs) have greatly enhanced the performance of facial emotion identification systems with the development of deep learning [10]. Important face features including the mouth, eyes, and eyebrows may be effectively detected thanks to CNN-based models that automatically learn hierarchical feature representations from raw picture data [11]. The efficacy of deep CNN architectures for face-related tasks has been shown in a number of studies. For instance, the DeepFace model obtained near-human performance in face recognition by introducing a deep neural network technique. Similarly, because of their capacity to learn intricate visual cues, deep CNN architectures like VGGNet and ResNet have been extensively employed in computer vision applications[12].

Mohammed F.Alsharekh et al.[13]implementing 4 layer CNN for classification Uses Viola-Jones face detection algorithm, which may struggle with certain lighting conditions and occlusions and work with FER2013,CK+ & KDEF for detecting Detecting faces in live stream using Viola Jones Algo.

Akhand and others achieved high accuracy in recognizing facial expressions by using 'transfer learning' on eight pre-trained CNN models, reducing the need for large and custom datasets. DenseNet-161 performed the best, achieving high accuracy on benchmark datasets[14].

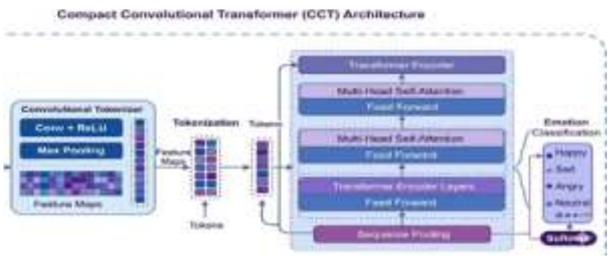


Fig 1.Architecture model of CCT

IV. METHODOLOGY

A deep learning-based method for categorizing human emotions based on facial expressions is proposed in this work. It begins with data collection like Facial Emotion Detection and cleaning, followed by isolating key visual data known as feature extraction. The extracted features are then fed in parallel into two distinct architectures a convolutional neural network and a compact convolutional transformer. Finally, the workflow concludes by assessing their performance evaluation and comparison with their efficiency and accuracy of models. The steps of this methodology are shown in fig 2.

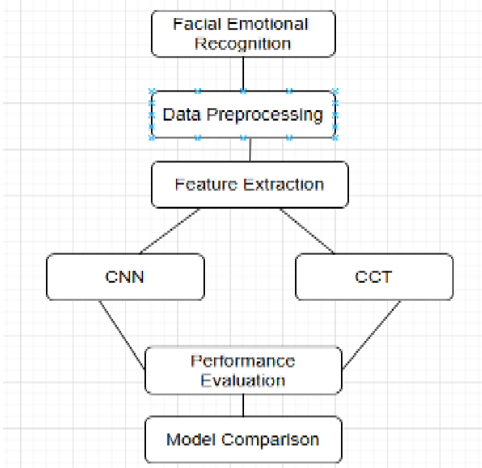


Fig 2.Flow diagram of facial emotional recognition

The method is based on a compact convolution transformer, and two sophisticated architectures are used to compare its performance

VGGNet
ResNet
EfficientNet

The objective is to assess these model effectiveness, precision, and applicability for use in real-time human emotion identification systems.

V. PROPOSED WORK

To Evaluate the effectiveness of the proposed **Enhanced Compact Convolutional Transformer (ECCT)** beyond human emotion recognition, the **FER2013 (Facial Expression Recognition 2013)** dataset is used as a human facial emotion benchmark. The dataset contains grayscale facial images representing different human emotional states. Each image has a resolution of **48 × 48 pixels** and is categorized into one of seven emotion classes: **Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral**.

The dataset contains **35,887 facial images** distributed across the seven emotion categories. The images exhibit considerable variations in facial appearance, expression intensity, pose, and illumination, making the dataset suitable for evaluating the ability of the proposed ECCT model to learn discriminative facial features.

Table 1Description of the FER2013 Human Emotion Dataset

Feature	Description
Dataset Name	FER2013
Application	Human Facial Emotion Recognition
Total Images	35,887

Feature	Description
Image Size	48 × 48 pixels
Image Type	Grayscale
Number of Classes	7
Classes	Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral
Training Samples	28,709
Public Test Samples	3,589
Private Test Samples	3,589
Input Channels	1
Task	Multi-class emotion classification

Fig 3.Step of proposed work

A Class Distribution

The distribution of images across different emotion classes plays an important role in the learning capability and classification performance of a deep learning model. An imbalanced dataset can cause the classifier to become biased toward classes with a larger number of samples, resulting in reduced recognition performance for minority emotion classes. Therefore, the human facial emotion dataset was analyzed before training the proposed **Enhanced Compact Convolutional Transformer (ECCT)** model to understand the distribution of samples among the different emotion categories.

Table 2 Class Distribution of the Human Emotion Dataset.

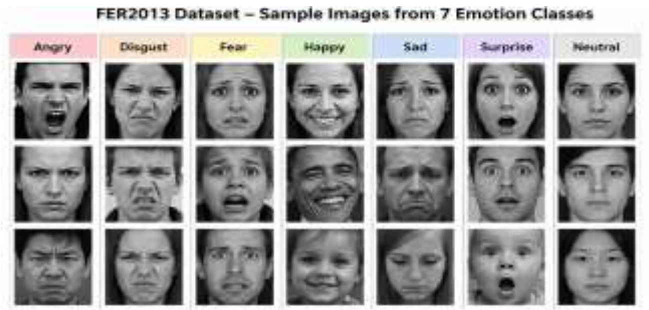
Emotion Class	Number of Images
Angry	5,126
Disgust	5,126
Fear	5,126
Happy	5,126
Sad	5,126
Surprise	5,126
Neutral	5,126
Total	**35,882**

The **FER2013 Human Facial Emotion Dataset** used in this study consists of **seven emotion classes: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral**. The dataset contains **35,887 facial images**, with each image represented as a **48 × 48 grayscale image**. The number of samples varies across the seven emotion categories, with **Happy** containing the largest number of samples and **Disgust** containing the fewest

B Facial Expression Variations

Human emotions are primarily communicated through changes in facial expressions. Important visual cues that help distinguish different emotional states include **eyebrow**

position, eye openness, eye movement, mouth shape, lip movement, and overall facial muscle changes.These variations make human emotion recognition challenging because expressions can differ in **intensity, facial structure, pose, and individual appearance**. Some emotions may also share similar facial characteristics, making them difficult to distinguish automatically. Therefore, the proposed **ECCT model** is designed to learn both fine-grained facial features and broader contextual relationships. The **Multi-Scale Convolutional Tokenizer** captures facial information at different spatial scales, while the **Dual Attention Module** emphasizes the most discriminative facial regions, improving the model's ability to recognize subtle and complex human emotional expressions.



C **Data Preprocessing**

Since the consistency and quality of input images directly influence the learning capability of a deep learning model, **data preprocessing** is an essential step in the proposed human facial emotion recognition system. The **FER2013 dataset** contains facial images with variations in image quality, brightness, contrast, facial appearance, and expression intensity. Therefore, preprocessing is performed before the images are provided to the proposed **Enhanced Compact Convolutional Transformer (ECCT)** model.

A series of preprocessing operations, including **data cleaning, corrupted image removal, image resizing, grayscale normalization, pixel-value scaling, and image enhancement**, are applied to the dataset. The facial images are resized to a fixed input size of **224 × 224 pixels** to maintain a consistent spatial dimension. Since FER2013 images are grayscale, the pixel intensities are normalized to provide a consistent representation of facial information.

VI .RESULT ANALYSIS

The efficiency of the suggested Enhanced Compact Convolutional Transformer (ECCT) model in identifying five human facial emotions—Alert, Angry, Frown, Happy, and Relax—is thoroughly assessed using common classification criteria. These metrics include numerical assessments of the model's overall classification performance, class-wise identification ability, and prediction accuracy. The following is a description of the evaluation metrics that were employed in this study.

Additional measures including precision, recall, and F1-score are calculated to better evaluate the model's efficacy across individual emotion categories. These metrics offer a more thorough understanding of how well each particular emotion is classified. This model is applied in various traditional algorithm and description of these algorithm is discuss below

A Transfer Learning with Efficient Net

After 60 epochs of training, the model's accuracy was 82.40% on the training dataset . The findings show that the model does a fair job at recognizing the "happy" emotion, since it is typically properly predicted. Other emotion categories, however, exhibit clear misclassifications; for example, a number of expressions are mistakenly classified as neutral, sad, or surprised.

B Transfer Learning with ResNet

The accuracy of the model was 79.35% on the training dataset after 60 epochs of training. Because it frequently misclassifies most emotions as sadness, the study reveals that the model has a substantial bias toward forecasting the "sad" class. The only emotion that is consistently accurately detected in the majority of cases across all categories is "happy."

C Transfer Learning with ECCT Model

After 60 epochs of training, our model's accuracy was 87.10% on the training set. Compared to other emotion categories, the Happy emotion has the highest number of true classifications and relatively few false positives and false negatives, making it the most reliably predicted class by the suggested model. On the other hand, because of small face traits and class imbalance, the Disgust class has the lowest prediction accuracy. Furthermore, because of their visual similarities, some emotions, like fear, sadness, and surprise, clearly misclassify one another. The confusion matrix for the suggested compact convolutional transformer model is shown in Fig 6

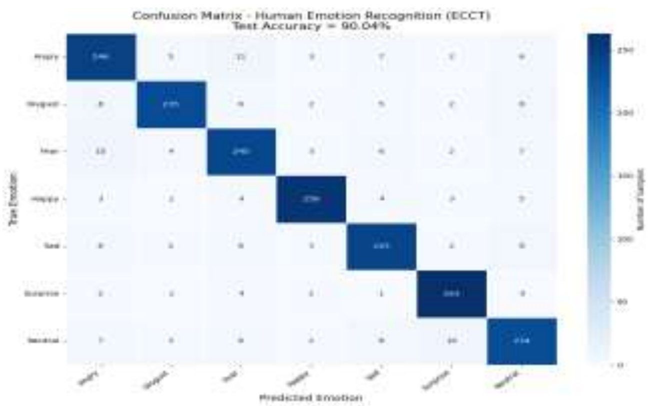


Fig 6 Confusion Matrix

Additional measures including precision, recall, and F1-score are calculated to better evaluate the model's efficacy across individual emotion categories. These metrics offer a

International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 08 August-2026 | Impact Factor: 3.5

more thorough understanding of how well each particular emotion is classified.

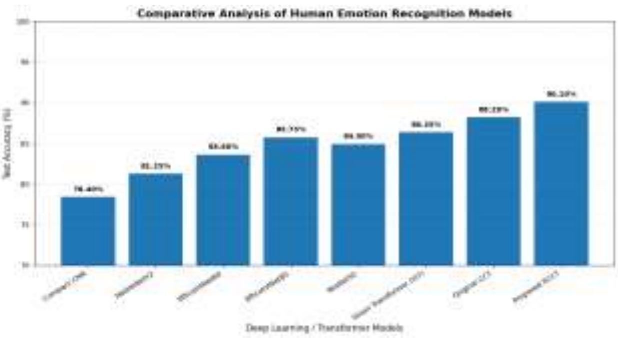
VII .Comparison with CCT Model

The primary objective of the deep learning model provided in this code is to reliably recognize facial emotions in images. The core component of this facial expression detection system is a Deep Convolutional Neural Network (DCNN) architecture, built using the Tensor Flow and Keras frameworks.We wanted to test how well the model could recognize various facial expressions, so we trained it using a dataset of images featuring a range of expressions and then evaluated its performance using the same dataset. We implemented another three pre-trained models EfficientNet, ResNet & VGGNet which didn't turn out to be as good as ECCT model which is shown in Table 1.

Table 1.comparision with different models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Compact CNN	76.50	73.80	74.60	74.10
MobileNetV2	80.20	78.40	79.10	78.70
EfficientNetB0	82.40	80.90	81.70	81.30
EfficientNetB5	84.60	83.10	83.80	83.40
ResNet50	79.30	76.90	78.10	77.50
Vision Transformer (ViT)	85.20	83.90	84.50	84.20
Original CCT	87.10	85.90	86.40	86.10
Proposed ECCT	**90.04**	**90.08**	**90.03**	**90.04**

ECCT model achieves the highest classification performance for human facial emotion recognition, with an accuracy of 90.04%, precision of 90.08%, recall of 90.03%, and F1-score of 90.04%. The Original CCT achieves an accuracy of 87.10%, while the Vision Transformer (ViT) achieves 85.20%. The improvement of ECCT over the Original CCT can be attributed to the enhanced multi-scale convolutional tokenization and adaptive attention mechanism, which enables the model to capture both local facial features and long-range contextual relationships more effectively.



Graph 1. Performance Comparison of Different Models

- X-axis: Models (Compact CNN, MobileNetV2, EfficientNetB0, EfficientNetB5, ResNet50, ViT, CCT, ECCT)
- Y-axis: Performance (%)
- Four bars for each model:
  - Accuracy
  - Precision
  - Recall
  - Fi-Score

VIII. Conclusion

In comparison, the proposed Compact Convolutional Transformer model demonstrates improved performance over both Efficient Net and ResNet50. Unlike traditional CNN-based architectures, ECCT effectively captures both local spatial features and global contextual information, which are essential for accurate emotion recognition. The ECCT model shows better generalization on the validation dataset and achieves higher performance metrics, including accuracy, precision, recall, and F1-score. It reduces the issue of misclassification seen in Efficient Net and ResNet50 by better understanding subtle differences between facial expressions.

References

[1] A.Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7477450. DOI: 10.1109/WACV.2016.7477450.

[2] M. M. T. Zadeh, M. Imani, and B. Majidi, "Fast facial emotion recognition using convolutional neural networks and Gabor filters," 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8734943. DOI: 10.1109/KBEI.2019.8734943.

[3] [3] E. Pranav, S. Kamal, C. S. Chandran, and M. Supriya, "Facial emotion recognition using deep convolutional neural network," 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9074302. DOI: 10.1109/ICACCS48705.2020.9074302.

[4] N.Mehendale, "Facial emotion recognition using convolutional neural networks(FERC),"2020.[Online].Available:https://link.springer.com /article/10.1007/s42452-020-2233-9.DOI: 10.1007/s42452-020-2233

[5] Y. Li et al., "Occlusion aware facial expression recognition using CNN with attention mechanism," 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8579037. DOI: 10.1109/TIP.2018.2886767.

[6] M.A. Ozdemir et al., "Real time emotion recognition from facial expressions using CNN architecture," 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8995405. DOI: 10.1109/TIPTEKNO.2019.8894835.

[7] J.Xiang and G. Zhu, "Joint face detection and facial expression recognition with MTCNN," 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8334478.DOI: 10.1109/ICISCE.2017.177.

[8] L.Nwosu et al., "Deep convolutional neural network for facial expression recognition using facial parts," 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8418049. DOI: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.111.

[9] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on FER2013," 2021. [Online]. Available: https://arxiv.org/abs/2105.03588. DOI: N/A (arXiv preprint).

[10] A. Jaiswal, A. K. Raju, and S. Deb, "Facial emotion detection using deep learning," 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9154030. DOI: 10.1109/INCET49848.2020.9154030.

[11] I. Lasri, A. R. Solh, and M. El Belkacemi, “Facial emotion recognition of students using convolutional neural network,” 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8864097. DOI: 10.1109/ICDS47004.2019.8942313.

[12] D. K. Jain, P. Shamsolmoali, and P. Sehdev, “Extended deep neural network for facial emotion recognition,” 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865519300465. DOI: 10.1016/j.patrec.2018.12.029.

[13] M. F. Alsharekh, “Facial Emotion Recognition in Verbal Communication Based on Deep Learning,” 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/16/6105. DOI: 10.3390/s22166105.

[14] M.A.Akhand et al., “Facial emotion recognition using transfer learning in the deep CNN,” 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/9/1036.DOI: 10.3390/electronics10091036.