

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

Sahar Afzal¹, Bharat Singh Ladhi², Deepti razdan³

^{1,2}*Dept of Computer Application, Eklavya Univercity, Damoh, M.P*

³*Dept of Computer Science and Engineering, Eklavya Univercity, Damoh, M.P*

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 compact convolution Transformer provide remarkable superior result for detecting a facial emotion and give best result as compare to convolution neural network. This research recognises 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 happy, sad, angry, fear, disgust, neutral, and surprise. This experiment is conducted in FER2013 dataset. The proposed model achieves an accuracy of 0.63 as compare to VGG16, VGG19, and Efficient Net.

Index Terms—Convolution Neural Network, Efficient Net, Compact Convolutional Transformer, Facial emotional recognition, deep neural network.

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

1. Develop a real time emotional recognition with hybrid model.
2. Design a compact transformer model for comparison with convolution models.
3. Experiment is conducted in FER2013 dataset to assess both models' performance.
4. 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 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].

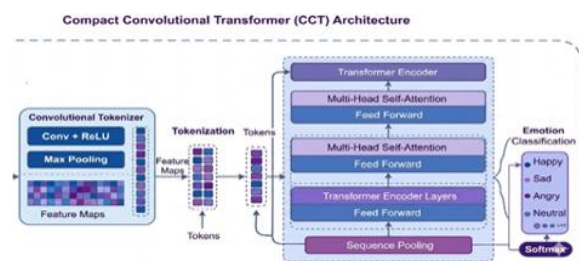


Fig 1. Architecture model of CCT

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.

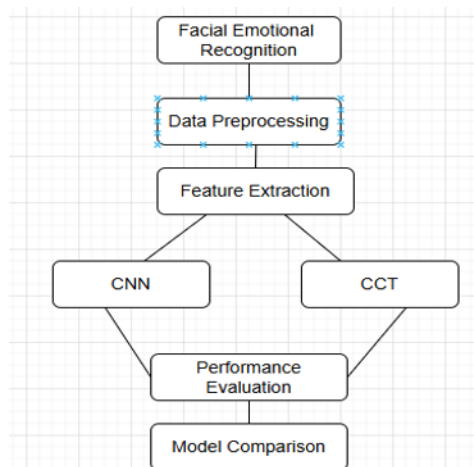


Fig 2. Flow diagram of facial emotional recognition

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

1. VGGNet
2. ResNet
3. EfficientNet

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

V. PROPOSED WORK

In this research, hybrid model which is the combination of CNN or Vision Transformer is used that can identify human emotions by looking at human face images. fig 3 shows the steps of performing of this research

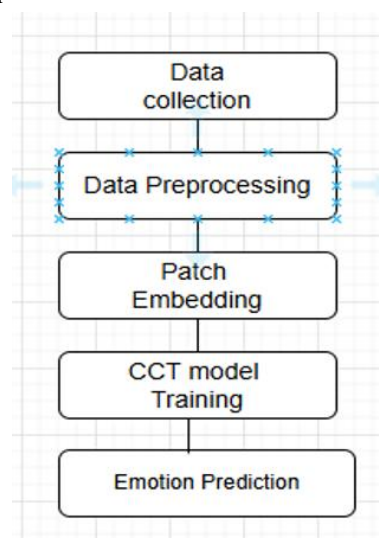


Fig 3. Step of proposed work

A Data Collection

The FER2013 Dataset is used in this work to recognize face emotions shown in Fig 4. This dataset comprises More than 35,887 facial photos labeled with seven main emotional categories.



Fig 4. Feature extraction before CCT

The photos range in size and resolution and depict people from various age groups, genders, and cultural pictures were collected from publicly accessible websites like Google pictures and Flickr in order to create the datasets. To confirm the dataset's trustworthiness and accuracy, a subset of the photos was manually verified to form a validation set. Furthermore, to prepare the datasets for efficient model training, preprocessing approaches were used, such as resizing the photos to a consistent dimension and converting them to gray scale format to reduce computational complexity.

B Data Preprocessing

To increase data quality and effective learning, datasets are preloaded before training the model. This phase is used to normalize the training process and minimizes computing complexity. To improve the model's robustness, several data augmentation approaches are applied to the training images. These changes broaden the datasets diversity and enable the algorithm to learn more generalized facial traits. During training, these procedures produce diverse variants of the same image, allowing the model to learn invariant features from various angles. Separate data loaders are developed for the training and validation datasets in order to train and evaluate models. Throughout the learning process these loaders provide batches of images together with their matching emotion label. The incoming photographs are downsized to a preset dimension and processed in batches of 128 images to allow for efficient processing. The processed data is then utilized to train

a compact convolutional transformer architecture. This model uses convolutional tokenization and transformer-based attention methods to capture both local and global facial features. The proposed approach seeks to increase the model's generalization ability and emotion identification accuracy in real world scenarios by using preprocessing and augmentation strategies.

C Model Selection

In this research one of the deep learning models is used for facial emotion recognition known as compact convolutional transformer which is a hybrid architecture that effectively learns both local and global representations from images by combining transformer-based attention mechanisms with convolutional operations. The FER2013 Dataset's 48×48 gray scale facial expression photos are used in Fig 1.4 to train the suggested model. The model architecture is made to efficiently extract emotional clues and significant face patterns from these pictures. The CCT model first uses a convolutional tokenizer to extract low-level spatial characteristics from the input image which is shown in Fig 5, in contrast to conventional CNN designs that solely depend on stacked convolution and pooling layers. The image is divided into smaller feature patches by this tokenizer, which then transforms them into tokens that may be processed by transformers. After that, these tokens go through a series of transformer encoder layers that use self-attention techniques to discover connections between various face regions.

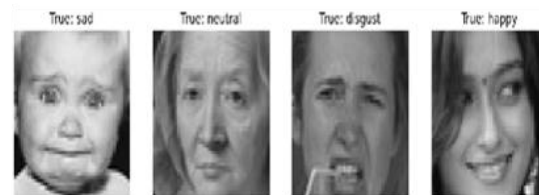


Fig 5. feature extraction after CCT

VI. RESULT ANALYSIS

The model is analyzed in this research utilizing the FER2013 dataset using a variety of performance measures, including precision and other relevant metrics. A different test dataset that was previously separated from the original dataset is used for evaluation. To test the trained model prediction power,

this test data is fed into it. The ratio of accurately predicted images to the total number of test samples is used to calculate the overall accuracy.

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 this algorithm is discuss below

A Transfer Learning with Efficient Net

After 50 epochs of training, the model's accuracy was 62.15% on the training dataset and 58.41% on the validation 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 59.40% on the training dataset and 54.67% on the validation dataset after 50 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 CCT Model

After 50 epochs of training, our model's accuracy was 63.18% on the training set and 55.00% on the validation 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

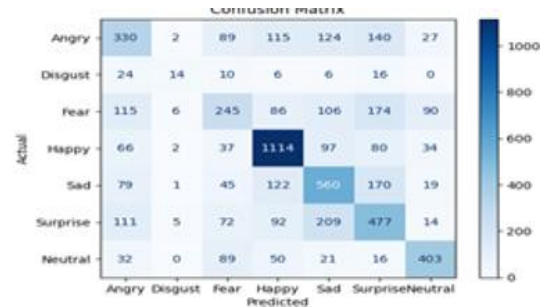


Fig 6 Confusion Matrix

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 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 CCT model which is shown in Table 1.

Table 1. Comparison with different models

Model	Training Accuracy	Validation Accuracy	Precision	Recall	F1-Score
EfficientNet	0.6215	0.5841	0.36	0.26	0.28
ResNet50	0.5941	0.5467	0.23	0.24	0.2
VGGNet	0.5031	0.5111	0.17	0.15	0.15
CCT (Proposed)	0.63	0.55	0.54	0.55	0.54

Therefore, CCT is the best proposed model. On the training data, the best proposed emotion recognition model achieved an accuracy of 63%, and on the validation data, it achieved an accuracy of 55%. The developed model can recognize emotions such as happiness, sadness, fear, anger, disgust, surprise, and

neutrality. Potential applications of this model include human-computer interaction, marketing research, and mental health assessment. Further research could be conducted to improve the model's accuracy and evaluate its performance in real-world scenarios.

VIII. CONCLUSION

In comparison, the proposed Compact Convolutional Transformer model demonstrates improved performance over both Efficient Net and ResNet50. Unlike traditional CNN-based architectures, CCT effectively captures both local spatial features and global contextual information, which are essential for accurate emotion recognition. The CCT 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] Mollahosseini A, Chan D, Mahoor MH. Going deeper in facial expression recognition using deep neural networks. 2016.
- [2] Zadeh MMT, Imani M, Majidi B. Fast facial emotion recognition using convolutional neural networks and Gabor filters. 2019.
- [3] Pranav E, Kamal S, Chandran CS, Supriya M. Facial emotion recognition using deep convolutional neural network. 2020.
- [4] Mehendale N. Facial emotion recognition using convolutional neural networks (FERC). 2020.
- [5] Li Y, et al. Occlusion aware facial expression recognition using CNN with attention mechanism. 2018.
- [6] Ozdemir MA, et al. Real time emotion recognition from facial expressions using CNN architecture. 2019.
- [7] Xiang J, Zhu G. Joint face detection and facial expression recognition with MTCNN. 2017.
- [8] Nwosu L, et al. Deep convolutional neural network for facial expression recognition using facial parts. 2017.
- [9] Khairuddin Y, Chen Z. Facial emotion recognition: State of the art performance on FER2013. 2021. arXiv preprint.
- [10] Jaiswal A, Raju AK, Deb S. Facial emotion detection using deep learning. 2020.
- [11] Lasri I, Solh AR, El Belkacemi M. Facial emotion recognition of students using convolutional neural network. 2019.
- [12] Jain DK, Shamsolmoali P, Sehdev P. Extended deep neural network for facial emotion recognition. 2019.
- [13] Alsharekh MF. Facial Emotion Recognition in Verbal Communication Based on Deep Learning. 2022.
- [14] Akhand MA, et al. Facial emotion recognition using transfer learning in the deep CNN. 2021.