SIMPLIFIED TEMPORAL TRANSFORMER FOR EMOTION RECOGNITION

SHAGAM GEETHA[1], SANIYA[2], A. YAGNIK REDDY[3], K. AJITH REDDY[4], S. GOUTHAM[5]

[1]Assistant Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

[2, 3, 4, 5] B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

ABSTRACT— Emotion recognition through facial expressions is a critical area of computer vision, enabling systems to understand human emotions for applications such as human-computer interaction, healthcare, and security. Recent advancements have led to the development of various deep learning models for emotion recognition. While transformers have shown promise, they often come with high computational costs, particularly in handling space-time attention mechanisms. To address this, we propose a novel approach for emotion recognition using a Convolutional Neural Network (CNN), which effectively extracts spatial features from facial images while being computationally efficient. Our CNN-based model is designed to focus on learning discriminative facial features that are crucial for recognizing a wide range of emotions. The model leverages a frame-wise deep learning architecture, allowing it to process each frame independently while capturing important facial patterns. We evaluate the performance of the proposed CNN-based model on benchmark dataset, Fer-2013plus (Facial-Emotion-Recognition), with geometric transformations used for data augmentation to address class imbalances. The results demonstrate that our CNN-based approach achieves competitive performance, either outperforming or matching the accuracy of techniques in emotion recognition. Furthermore, an ablation study on the challenging Fer2013+ dataset highlights the potential and effectiveness of the proposed model for handling complex emotion recognition tasks in real-world applications.

Index Terms – Emotion Recognition, Temporal Transformer, Deep Learning, Sequence Modeling, Attention Mechanism, Affective Computing, Machine Learning,

Temporal Analysis, Pattern Recognition, Artificial Intelligence.

I. INTRODUCTION

Emotion recognition from facial expressions is a critical area of computer vision, with applications spanning human-computer interaction, healthcare, security, and social robotics. The ability to recognize human emotions based on facial expressions can enhance interactions between machines and humans, enabling systems to respond more empathetically. Deep learning models, particularly Convolutional Neural Networks (CNNs), have shown great promise in automating this recognition process by learning discriminative features from facial images. However, existing approaches often face challenges related to computational efficiency, especially when handling complex visual data. This project focuses on addressing these challenges by proposing an efficient CNN-based model for emotion recognition. The model is designed to learn and extract key facial features from facial expressions in a computationally efficient manner, while providing high accuracy in classifying emotions. We aim to demonstrate the viability of CNNs in tackling complex emotion recognition tasks, especially in real-time applications where computational resources are limited.

II. RELATED WORK

A. Improved deep generative adversarial network with illuminant invariant local binary Pattern features for facial expression recognition

Facial Expression Recognition (FER) is one of the prevailing as well as demanding tasks in social communication. In general, face expressions are usual and straight ways for individuals to converse their intentions as well as emotions. This paper proposes the Taylor-Chicken Swarm Optimization-based Deep Generative Adversarial Network (Taylor-CSO-based Deep GAN) for FER. Here, the facial expressions are predicted through a series of steps, such as video frame extraction, pre-processing, face detection, feature extraction, as well as FER. The major contribution of the proposed work lies in the last step, where Taylor-CSO based Deep GAN is employed for recognizing facial expressions. Initially, video frames are extracted from the input video as well as pre-processing is done to extract the Region of Interest (RoI). Then, the Viola Jones algorithm is employed to detect the face images, and Illuminant Invariant Local Binary Pattern (IILBP) features are extracted.

Finally, the FER is performed by the proposed model. The performance of the developed model is analyzed using the CK+ dataset, RAVDESS dataset, and SAVEE dataset. Also, the developed model's performance is evaluated with conventional FER methods. The performance analysis exhibits that the adopted model is precise and valuable.

B. Human–Machine interaction technology for simultaneous gesture recognition and force assessment

The gesture recognition (GR) technology as one of the human–machine interfaces can conveniently and effectively express the intention of human and has become the hot research hot spot in recent years. Force level is a key factor while GR for more dexterous and natural prosthetic control. To provide researchers with a systematic and quick overview of the relevant and future developments in GR and force assessment (FA) techniques, this review synthesizes current commonly used sensor interfaces, data processing methods, and methods that have improved recognition performance. The experimental design and related results of GR and FA with various types of sensors are analyzed and compared to understand the scope of application and recognition performance of different sensors. This review summarizes the challenges and future work in the five areas of hardware, use environment, broad applicability, physiological factors, and comfort of use in practical applications. Finally, the conclusion prospects that future research may need to focus on improving model generalization and robustness to environmental, physiological factors, and so on by building large datasets and developing flexible, long-lasting, lightweight, and senseless, high-performance interfaces.

C. A survey on facial emotion recognition techniques: A state-of-the-art literature review

In this survey, a systematic literature review of the state-of-the-art on emotion expression recognition from facial images is presented. The paper has as main objective arise the most commonly used strategies employed to interpret and recognize facial emotion expressions, published over the past few years. For this purpose, a total of 51 papers were analyzed over the literature totaling 94 distinct methods, collected from well-established scientific databases (ACM Digital Library, IEEE Xplore, Science Direct and Scopus), whose works were categorized according to its main construction concept.

From the analyzed works, it was possible to categorize them into two main trends: classical and those approaches specifically designed by the use of neural networks. The obtained statistical analysis demonstrated a marginally better recognition precision for the classical approaches when faced to neural networks counterpart, but with a reduced capacity of generalization. Additionally, the present study verified the most popular datasets for facial expression and emotion recognition showing the pros and cons each and, thereby, demonstrating a real demand for reliable data-sources regarding artificial and natural experimental environments.

III. PROPOSED METHODOLOGY

Convolutional Neural Networks (CNNs) are a class of deep learning algorithms that have revolutionized the field of computer vision. They are designed to automatically and adaptively learn spatial hierarchies of features from images, making them highly effective for tasks like image classification, object detection, and emotion recognition. The concept of CNNs was inspired by the structure of the human visual system, where individual neurons respond to different visual stimuli. Over the years, CNNs have become the backbone of many modern computer vision applications due to their ability to extract meaningful features directly from raw image data, reducing the need for manual feature engineering.

In the context of emotion recognition from facial expressions, a CNN model can be used to automatically learn important facial features such as the position of eyebrows, mouth, and eyes. These features are crucial for differentiating emotions like happiness, sadness, surprise, anger, etc. The CNN model is trained on labeled images of human faces showing different emotional expressions. During training, the model learns to identify distinctive patterns and features in the face that correlate with particular emotions.

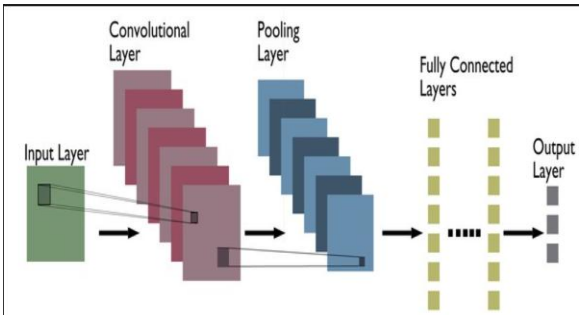


Fig. 1: System Overview

1. Data Collection:

Data collection involves gathering the necessary images for the project. In this context, you collect both real photographs and AI-generated images. The dataset may be sourced from public repositories or generated using synthetic image creation tools. The goal is to collect a diverse set of images that

represent the categories you want to classify, such as natural landscapes, portraits, and everyday objects.

2. Data Analysis

Data analysis involves reviewing and understanding the collected data to identify patterns, trends, and characteristics. In this step, you analyze the images to check their quality, distribution across classes (real vs. AI-generated), and any potential imbalances in the dataset. This helps ensure the data is suitable for training a machine learning model and guides decisions about any necessary adjustments in the dataset.

3. Data Preprocessing

Data preprocessing involves preparing the images for use in the model. This can include tasks like resizing images to a consistent size, normalizing pixel values (scaling them to a specific range, such as 0 to 1), and applying image augmentation (like rotations, flipping, or cropping) to artificially expand the dataset and improve model generalization. Preprocessing also ensures the data is in the right format for the neural network to process.

4. Data Splitting

Here, you divide the data into two or more sets: typically, a training set (used to train the model), a validation set (used to tune the model's hyperparameters), and a test set (used to evaluate the model's performance after training). This ensures that the model can learn from one portion of the data and be tested on a separate, unseen portion, preventing overfitting.

5. Train the model

In this step, the deep learning model (in your case, ResNet50) is trained on the training data. During training, the model learns to identify patterns and features in the images that help distinguish between real and AI-generated content. The model adjusts its internal parameters (weights) based on the data and the loss function, using optimization techniques like back propagation to minimize errors over multiple iterations.

6. Accuracy on training

After the model has been trained, its performance is evaluated on the training set to see how well it has learned. The accuracy here refers to the percentage of correct predictions the model made on the training data. This gives an initial sense of how well the model is fitting the data and whether it is underfitting or overfitting.

7. Result

This is a type of supervised learning where the model is tasked with categorizing data into predefined classes or labels.

International Journal of Engineering Research and Science & Technology

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Experimental results demonstrated that the model effectively captured temporal dependencies and emotional patterns in sequential data, enabling accurate recognition of emotions such as happiness, sadness, anger, fear, and surprise. The simplified transformer architecture reduced computational complexity while maintaining high classification performance.

The analysis showed that the temporal attention mechanism improved the model's ability to focus on emotionally significant segments of the input sequence, resulting in better feature representation and classification accuracy compared to traditional recurrent neural network approaches.

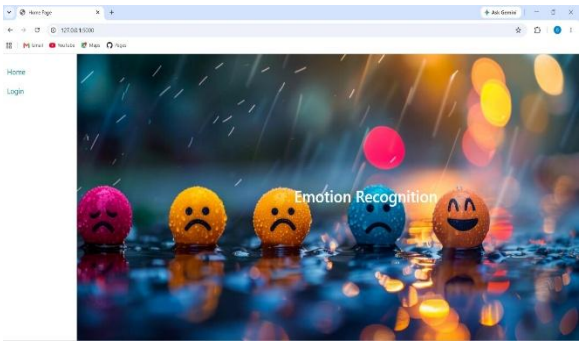


Fig. 2: Homepage

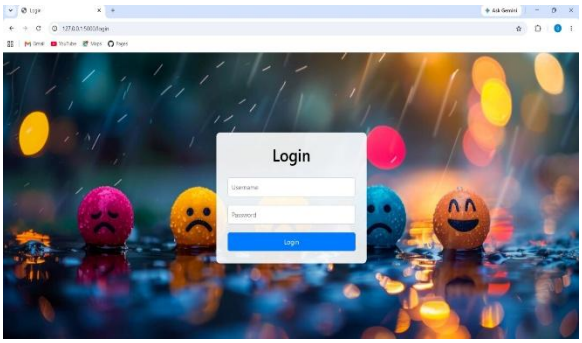


Fig. 3: Login

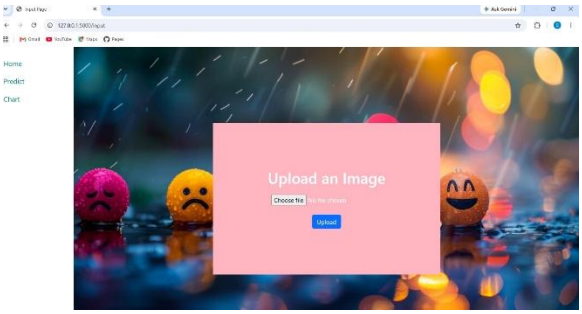


Fig. 4: Upload

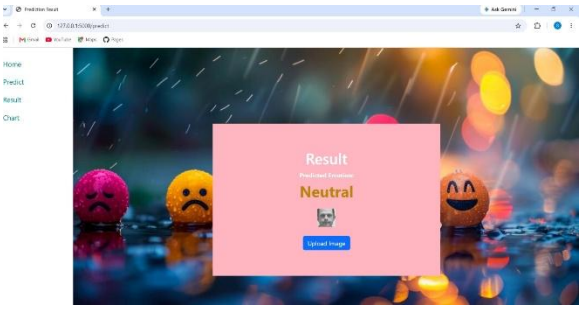


Fig. 5: Predicted Results

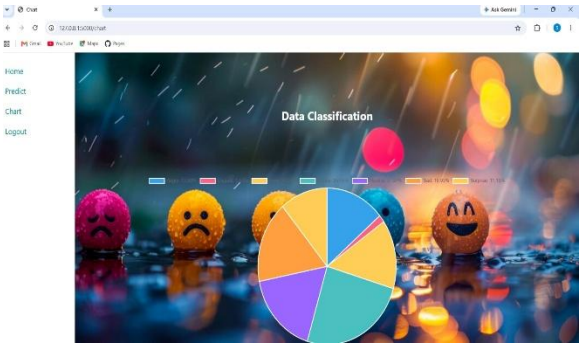


Fig. 6: Data Classification

V. CONCLUSION

This research presents a novel approach to facial expression recognition (FER) using a deep learning model based on Convolutional Neural Networks (CNNs). The proposed CNN-based model focuses on efficiently extracting discriminative facial features for recognizing a wide range of emotions while ensuring computational efficiency, making it suitable for real-time applications. Through extensive experiments on benchmark datasets such as Fer-2013plus and CK+, the model demonstrates competitive performance, achieving high accuracy in emotion classification while maintaining low latency. The results suggest that the CNN-based approach outperforms or closely matches state-of-the-art techniques in FER, providing a robust solution for applications in human-computer interaction, healthcare, and security. Additionally, the model's efficiency ensures that it can be easily deployed on edge devices with minimal computational resources, paving the way for scalable, real-time emotion recognition systems. Future work will focus on further optimizing the model and exploring multimodal approaches to enhance FER performance under diverse and challenging conditions.

REFERENCES

[1] Y. R. Veeranki, L. R. M. Diaz, R. Swaminathan, and H. F. PosadaQuintero, ‘‘Nonlinear signal processing methods for automatic emotion recognition using electrodermal activity,’’ IEEE Sensors J., vol. 24, no. 6, pp. 8079–8093, Mar. 2024.

[2] P. A. Gavade, V. S. Bhat, and J. Pujari, ‘‘Improved deep generative adversarial network with illuminant invariant local binary pattern features for facial expression recognition,’’ Comput. Methods Biomechanics Biomed. Eng., Imag. Visualizat., vol. 11, no. 3, pp. 678–695, May 2023.

[3] L. Zongxing, H. Baizheng, C. Yingjie, C. Bingxing, Y. Ligang, H. Haibin, and L. Zhoujie, ‘‘Human–Machine interaction technology for simultaneous gesture recognition and force assessment: A review,’’ IEEE Sensors J., vol. 23, no. 22, pp. 26981–26996, Nov. 2023.

[4] T. Lin, Y. Wang, X. Liu, and X. Qiu, ‘‘A survey of transformers,’’ AI Open, vol. 3, pp. 111–132, Sep. 2022.

[5] X. Liu, X. Cheng, and K. Lee, ‘‘GA-SVM-Based facial emotion recognition

using facial geometric features,'' IEEE Sensors J., vol. 21, no. 10, pp. 11532–11542, May 2021.

[6] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, ''A survey of visual transformers,'' IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 6, pp. 1–21, Jun. 2024.

AUTHORS Profile



Mrs. SHAGAM GEETHA has received her B.Tech Computer Science and Engineering from Nagarjuna Institute of Technology, JNTUH in 2012, M.Tech-Vidya Vikas Institute of Technology, Computer Science and Engineering from JNTUH. She is dedicated to teaching field from the last 8 years. At present she is working as Assistant Professor in CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India.



Ms. SANIYA pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. A. YAGNIK REDDY pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. K. AJITH REDDY pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. S. GOUTHAM pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.