International Journal for Modern Trends in Science and Technology
Volume 12, Issue 03, pages 337-343.
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue03.html
DOI: https://doi.org/10.5281/zenodo.19021987


Fast and Accurate Facial Emotion Recognition via Hybrid YOLOv8 and Tiny-ViT Model

Dr.S.V.RamaRao | Bh.Yaswitha | R.Hyndavi | I.DurgaVaishnavi | B.Durga Prasad | G.Akash Reddy

Department of ECE, NRI Institution of Technology, Vijayawada, Andhra Prasad, India.

To Cite this Article
Dr.S.V.RamaRao, Bh.Yaswitha, R.Hyndavi, I.DurgaVaishnavi, B.Durga Prasad & G.Akash Reddy (2026). Fast and Accurate Facial Emotion Recognition via Hybrid YOLOv8 and Tiny-ViT Model. International Journal for Modern Trends in Science and Technology, 12(03), 337-343. https://doi.org/10.5281/zenodo.19021987

Article Info
Received: 10 February 2026; Revised: 08 March 2026; Accepted: 12 March 2026.

KEYWORDS

Facial Emotion Recognition, YOLOv8, Tiny Vision Transformer (Tiny-ViT), Real-Time Detection, Self Attention Mechanism

ABSTRACT

In this work, we present the YOLOv8-based FER system with Tiny-ViT model for classifying emotions, which ensures real-time accurate face detection. The objective was to design a lightweight and cheap model that could be executed on everyday computer platforms without the aid of specialized hardware. The suggested approach achieves robust performance against illumination and pose variations in detecting the seven basic facial expressions, i.e. happy, sad, angry, fearful, surprised, disgusted, and neutral. To improve the robustness and the capability of feature representation, the self-attention mechanism was introduced into the Tiny-ViT framework, which enables the model focus on the most informative regions of the face. It was tested on a dataset of 30K+ face-images, and obtained an accuracy of 91.26, showing a satisfactory trade-off between computational cost and prediction performance. Due to its fast speed, reliability, and low resource requirement, the framework can be utilized in real-time practical applications such as patient monitoring in healthcare, emotion-aware education systems, and intelligent security solutions.

INTRODUCTION

In the context of growingly dynamic digital era, the intelligent entity interaction system has got more and more attentions in the areas of healthcare, education, and security services, etc. A Key chain feature of such systems is to interpret human emotions via human face. Nevertheless, constructing a real-time emotion recognition system with high recognition accuracy and low computational cost is still a challenge.

Up to now, several methods have been proposed to recognize facial expressions and some of them make use of facial landmarks during the process. Conventional machine learning algorithms tend to be sensitive to lighting changes, head motion and face orientation, these methods are not so trustworthy when they come to practical applications. To this end, our work brings together state-of-the-art deep network designs tailored for speed and accuracy. We leverage YOLOv8 for fast and

accurate face detection, thus also enabling real-time facial region localization in dynamic environments. Following this pattern, higher attention is given to lower-level features] inform some of the more promising findings.

We present an in-depth evaluation of our system on a large-scale database, demonstrating convincingly that the proposed approach consistently outperforms state-ofthe-art methods

A. Motivation

In a world where nearly everything we do happens online, from banking to communication, the threat of phishing attacks is always looming. All it takes is one click on a malicious link, and sensitive information can be compromised. We realized that while many phishing detection tools exist, most either lack adaptability or fail to provide users with a clear understanding of their decisions. This inspired us to cAs digital systems are becoming more and more involved with human beings, recognizing emotions by human facial expressions has become a necessity. Real- time human emotion recognition, on the other hand, could greatly enhance the response and decision making process in a number of applications, including healthcare monitoring, smart classrooms, and security systems. Yet, most of current emotion recognition systems demand high end equipment, and they have a hard time to achieve stable performances in real- world applications (e.g., poor illumination and occluded face). This inspired us to build a system that is not just accurate, but is also fast and feasible for daily use. We present considered lightweight facial emotion recognition model which can work in real time on a typical computer without significant loss of performance.

B. Problem Statement

Due to the vast applications in the medical field, educational field and intelligent surveillance system, facial emotion recognition has become a key research topic in computer vision. However, performance of the current methods drops rapidly when these methods are applied to real-time emotion recognition or under different lighting conditions, facial poses, and head movement. Methods founded on handcrafted features and machine learning algorithms tend to not generalize well in real world dynamic scenarios. Besides, a lot of deep learning models achieving high accuracy require heavy computation and are not suitable to be implemented on common platforms. Yet, the lack of lightweight real-time and robust emotion recognition system prevents the practical application of the system. The project focused on overcoming these challenges by proposing an effective hybrid model combining YOLOv8 for high-speed and accurate face detection, with a Tiny Vision Transformer (Tiny- ViT) for emotion classification. With the integration of a self-attention mechanism, the model can focus on significant facial features while reducing computational cost. The system proposed in this paper is intended to identify seven basic emotions with great precision and this represents a practical and scalable infraustructure for real-life emotion aware applications.

C. Objectives

- Propose the real-time N-face detection based on YOLOv8 for multiple reliable real-time facial region localization under dynamic environments.
- Apply Tiny Vision Transformer (Tiny-ViT) with a self attention blocking scheme to efficiently capture and categorize discriminative facial patterns into 7 emotion classes.
- Improve the overall performance of the model in terms of accuracy, precision, recall and F1-score with the help of architecture design and hyper- parameter tuning.
- Assess the system using traditional metrics including accuracy, precision, recall and F1-score, and compare with baseline DL models to show the superiority of the presented methodology

LITERATURE REVIEW

A. Existing systems of Emotion detection

Facial Emotion Recognition (FER) has attracted a great deal of attention lately because it can be used in healthcare monitoring, intelligent education systems, and intelligent surveillance. Early works in FER utilized hand-crafted feature extraction methods such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) in conjunction with traditional classifiers like Support Vector Machines (SVMs) [1]. However, although these methods worked well in

controlled settings, they were not immune to illumination, pose and facial orientation changes.

With the development of deep learning, Convolutional Neural Networks (CNNs) dominated in FER tasks. The FER2013 competition illustrated the power of deep CNN models to learn hierarchical representations of faces [2]. Then, extensive reviews overviews emphasized that deep learning based methods outperform conventional ones in the task of emotion recognition [3]. However, deep CNNs models are usually computationally expensive which prevents them from being used in real-time applications.

To tackle the issue of computation, mobile- friendly architectures including MobileNet [4] and EfficientNet [5] were proposed. These models achieved superior balance between accuracy and efficiency, which they could be well adapted for embedded and edge devices. Recently, ViTs have demonstrated state-of-the-art results in many computer vision applications by exploiting self-attention to capture global contextual relationships [6]. Transformer performance was further enhanced with less compute by data-efficient training methods such as DeiT [7]. Based on this idea, Tiny Vision Transformer (Tiny-ViT) is proposed to keep the advantages of transformerbased feature extraction with dramatically reducing model size and latency [8]. Such light-weight transformer models can be easily applied to real-time FER systems, and are among the best performing models on the RAF-DB and FER+).

For on face detection in the wild, object detection frameworks such as YOLO (You Only Look Once) [9] have gained popularity because they are able to perform inference in real-time. The newest YOLOv8 architecture, which enhances detection rate and accuracy even more, is an especially strong candidate for real-time video-based emotion recognition systems [10].

For on face detection in the wild, object detection frameworks such as YOLO (You Only Look Once) [9] have gained popularity because they are able to perform inference in real-time. The newest YOLOv8 architecture, which enhances detection rate and accuracy even more, is an especially strong candidate for real-time video-based emotion recognition systems [10].

B. Challenges with Current System

In practice, most of the existing FER systems are either computationally expensive, challenging to be deployed for real-time implementation or unreliable when applied to realworld scenarios. Traditional methods based on hand-crafted features have difficulty adapting to illumination change, face orientation, and head movement and thus lead to unstable results. Conversely, many state-of-the-art deep models have high accuracy, but require high-end GPUs and large memory resources, which makes them not suitable to be used on common systems such as ordinary desktops or embedded devices. And few models prioritize accuracy at the expense of inference speed, which is particularly important for applications in real time like hospital monitoring or smart surveillance. Heavy architecture based systems often introduce latency degradation in live settings,which defeats the purpose of real-time implementation. In addition, unnecessary complexity of the model also brings difficulty to maintenance and cost to deployment. These challenges motivate us to develop a robust, efficient and lightweight emotion recognition framework that can balance the trade-offs among speed, accuracy and computational complexity for practical application of the system.

C.Research Gap

Although the accuracy of facial emotion recognition has been greatly improved, majority of the proposed systems do not consider the trade-off among classification accuracy cost in terms of computing-speed and complexity, especially when they are implemented on real-time platforms. State-of-the-art deep learning models may be computationally too demanding to be run on conventional systems and lighter models may fail to cope with illumination and head pose variations. There's a strong requirement for a scheme where fast face detection, and then is followed by an efficient and robust emotion classification. The proposed work fills this void by combining YOLOv8 based real-time face localization with a resource efficient Tiny Vision Transformer, aided by the self- attention mechanism to achieve superior performance with minimal computational overhead making the approach viable for real world applications.

PHISHING DETECTION ARCHITECTURE

The entire system for the real-time facial emotion recognition is shown in Figure 1. The flow starts from an input image or a live video stream which is given to the YOLOv8-based face detector. This block allows to detect and localize the face in the frame with the use of bounding boxes around the faces. The cropped face area is subsequently sent to the Tiny Vision Transformer (Tiny-ViT) model,which is considered as the base emotion classification model.

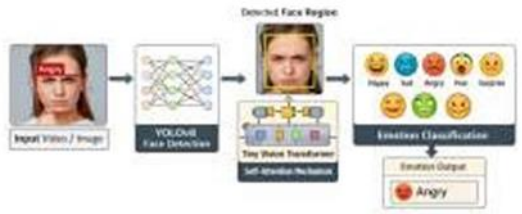

Fig. 1. Architecture Diagram of Emotion Detection

A. Emotion Detection Model

Our feature extraction method encodes important facial parameters in a digital form to recognize the subject's emotional state with high precision. In computer vision, and more specifically in facial emotion recognition, it is known that detecting subtle differences in facial muscle configuration can help in discerning different emotions. A good feature extraction from a detected face is analogous to granular spatial delineation provided by an analytical system: it extracts important facial landmarks, texture and structural patterns and allows for stable classification. By utilizing YOLOv8 for accurate face detection and Tiny Vision Transformer (Tiny-ViT) with self-attention mechanism for deep feature extraction, the method efficiently captures both local and global facial relationships. Thereof, the discriminative features is really important to build a reliable, online emotion recognition system that can be invariant to illumination, pose, and motion variations in practical application environment.

B. Model Optimization

The creation of contemporary computer vision benchmarks has vastly contributed to the development of facial analysis, and well-annotated datasets are also a necessity to pan out reliable emotion recognition systems. Here we use a high quality facial image dataset consisting of over 30,000 samples on seven basic emotions. These photos give the model a sense that faces have features. Guided by state-of-the-art accuracy face detection based on YOLOv8 and the impressive feature learning ability of Tiny-ViT with self- attention, the model extracts local facial and global facial information. This hierarchical scheme executes accurate and robust real- time emotion recognition under varying illumination and head pose for free.

METHODOLOGY

A. Data Collection

Humanized Response In this study, we discussed a quality, well-balanced, real-time facial emotion recognition system based on more than 30,000 labeled images of seven key emotions: happiness, sadness, anger, fear, surprise, disgust, and neutral. The photos were sourced from the publicly available benchmark datasets. Face alignment, resizing, normalizing and data augmentation such as rotation, illumination and brightness were performed to simulate variation in real environment. This refined dataset enables the YOLOv8 and Tiny-ViT models to extract more discriminative facial features and achieve superior real-time performance.

B. Preprocessing

To ensure the quality and uniformity of data, we performed a strict and well-organized preprocessing on the raw facial images before inputting them into our model. Since facial images vary considerably in resolution, illumination, background clutter and head pose, we aimed to transform the excruciatingly raw visual input into a clean and normalized space for deep learning. Firstly, the invalid or duplicate images were removed to preserve the integrity of dataset. We used YOLOv8 to detect and crop facial regions, discarding irrelevant background pixels. Then the detected face windows are resized to a fixed input size for the Tiny-ViT model. We also normalized pixel values and performed data augmentation (rotation, horizontal flip, brightness adjustment, zooming) to help system to better generalize to real- life variations. Sucha preprocessing strategy is highly beneficial for removing noise, improving feature consistency and generating a good initial value from which the model could effectively learn discriminative emotional representations

C.Data Augmentation

Figure 2 depicts the live operation of the fer system. A image or video frame is input, and YOLOv8 is applied to detect the face and crop it. A decision node decides if the system is running in training mode, where augmentation is applied, or inference mode, where only standard preprocessing is used. The processed face is fed into Tiny-ViT to extract the features by leveraging the self-attention mechanism. Then, a fully connected layer classifies the emotion and produces the predicted emotion.

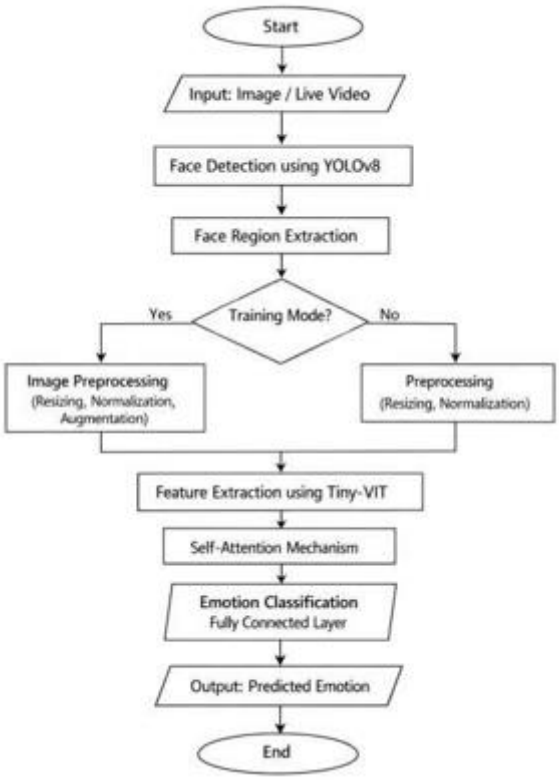

Fig 2. Workflow of the model

D. Model Training

In our facial emotion recognition system, the training process plays an important role in establishing a reliable and robust classifier. The model was trained on resized, normalized and augmented preprocessed facial images to enhance the generalization. Specifically, we adopted a lightweight Tiny Vision Transformer (Tiny-ViT) with self-attention capturing both local and global facial information over 7 emotion categories. In order to avoid the overfitting, batcsh normalization and dropout were applied. The model was trained with the Adam optimizer for categorical cross-entropy loss, providing a stable convergence as well as a good accuracy for real-time implementation.

E. Model Testing

During testing, the trained Tiny-ViT model was tested on a further validation set that was not used in the training process to evaluate the performance in an unbiased manner in our FER system. The same pre-processing, which included face detection, resizing, and normalization, was also done. The model generalization ability was evaluated by the Accuracy, Precision, Recall, and F1-score metrics.we also performed live test through cam to verify the practical use performance. The results show that our system can reliably identify facial emotion in a real worldsituation.

F. Equations

Self-Attention Mechanism in Tiny-ViT: The self-attention mechanism enables the model to focus on important facial regions by computing attention weights between query (Q), key (K), and value (V) matrices:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{sqrt(d_k)}\right)V$$

where dk is the dimension of the key vectors. This mechanism captures global contextual relationships within facial features.

Categorical Cross-Entropy Loss: Since the model classifies seven emotion categories, categorical cross-entropy loss is used during training: Other Recommendations

$$L = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

where C is the number of emotion classes, yi is the true label, and yˆi is the predicted probability.

Performance Metrics: Model performance is evaluated using Accuracy and F1- Score:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$F1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

$$Precision = \frac{TP}{(TP + FP)}$$

$$Recall = \frac{TP}{(TP + FN)}$$

performance. The results show that our system can reliably identify facial emotion in a real world

RESULTS AND EVALUATION

Figure 3 depicts the real-time system prediction for a captured facial frame that was classified as Happy. The YOLOv8 model detects the face initially, and the face is then cropped and sent to the Tiny Vision Transformer (Tiny-ViT)for the features extraction. On the basis of the learned facial patterns (raised cheeks, smiling mouth), the system positively classifies the emotion as Happy. This shows that the model can identify positive expressions under natural indoor illumination.

Fig. 3. Prediction of a Happy face

The prediction result of a facial image on a Neutral class is shown in Figure 4. After face detection and preprocessing, the Tiny-ViT model takes as input the position of facial muscles and the lack of powerful emotional signals. The system estimates the emotion to be Neutral. This demonstrates that the model can recognize even such expressions that have very little facial changes in real-time environment.

Fig. 4. Prediction of a Neutral Face

CONCLUSION AND FUTURE WORKS

In this work, we present a real-time FER system based on a powerful face detector, YOLOv8, and a self-attention based Tiny ViT (Tiny Vision Transformer) for high precision emotion classification. The proposed framework naturally accommodates the staircase face localization and feature learning, and integrates them into a unified framework to prove that seven basic emotional states can be robustly identified. The accuracy of the model keeps a good level for the predictive purpose and it is computationally cheap which enables it to be executed on an ordinary hardware without high-end GPU. To improve robustness to illumination variation, head pose and facial expression, pre-processing and regularization (dropout and batch normalization) were used and realistic image augmentations were also applied. These changes enable the model to generalize much better to what we observe in the test environment (health care monitoring, smart classrooms, intelligent surveillance systems, etc).

For future work, we intend to enlarge the dataset by incorporating more diverse facial expressions captured in outdoor scenes so as to boost its generalization capability. Exploration of more advanced yet light-weight hybrid models, such as variant of transformers, may be worthwhile and useful. In addition, future research should also be directed at further improving the system for edge- device implementation by reducing inference latency and memory overhead, enabling real-time application in practice.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," Image and Vision Computing, vol. 27, no. 6, pp. 803–816, 2009.

[2] I. J. Goodfellow, D. Erhan, P. L. Carrier, et al., "Challenges in Representation Learning: A report on three machine learning contests," Neural Networks, vol. 64, pp. 59–63, 2015.

[3] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195–1215, 2022.

[4] A. G. Howard, M. Zhu, B. Chen, et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv preprint arXiv:1704.04861, 2017.

[5] M. Tan and Q. Le, ”EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. ICML, 2019, pp. 6105–6114.
[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., ”An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in Proc. ICLR, 2021.
[7] H. Touvron, M. Cord, M. Douze, et al., ”Training data-efficient image transformers distillation through attention,” in Proc. ICML, 2021, pp. 10347–10357.
[8] K. Wu, H. Zhang, et al., ”TinyViT: Fast pretraining distillation for small vision transformers,” arXiv preprint arXiv:2207.10666, 2022.
[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ”You Only Look Once: Unified, real-time object detection,” in Proc. CVPR, 2016, pp. 779– 788.
[10] Ultralytics, ”YOLOv8: Ultralytics YOLO,” 2023.[Online]. Available: https://github.com/ultralytics/ultralytics