IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131114

Real-Time Crowd Emotion Detector Using YOLO and FER

Dr. M. P. Borawake1, Tanuja Dattatraya Jagtap[2], Samarendra Meher[3], Suhani Khandu Pawar[4], Shaili Balaji Tudme[5]

HOD Computer Engg, PDEA's College of Engineering, Manjari (BK), Pune, Maharashtra, India[1]

Department of Computer Engineering, PDEA's College of Engineering Manjari (BK), Pune, Maharashtra, India[2-5]

Abstract: Facial expression analysis is an active area of computer vision, but analysing a crowded scene presents practical challenges: several faces may appear in one frame, faces may be partially occluded or small, and illumination and head pose can vary. This paper presents the design of a Real-Time Crowd Emotion Detector that combines YOLO-based face detection with Facial Emotion Recognition (FER). The proposed pipeline accepts webcam or video input, extracts frames, detects multiple faces, crops the detected regions, predicts an emotion for each face, and aggregates predictions into a crowd-level emotion distribution and dominant category. The literature reviewed includes YOLO-based facial expression detection, video-based YOLO and Vision Transformer classification, FER-YOLO-Mamba, EmoNeXt, and OpenFace 3.0. These works motivate efficient detection, expressive feature learning, and attention to practical robustness. The current project is at the system-design and implementation stage; therefore, this paper reports the proposed architecture and evaluation plan rather than measured accuracy or speed. Evaluation is planned using labelled images and videos, with emotion-classification metrics, face-detection measures where labels are available, processing frames per second (FPS), and qualitative analysis of difficult scenes. The system is intended as an expression-analysis aid and does not claim to determine a person's internal emotional state with certainty.

Index Terms: crowd analysis, facial emotion recognition, YOLO, computer vision, real-time detection, emotion distribution

I. INTRODUCTION

Facial expressions provide visible cues that people use during social interaction. Computer-vision systems can analyse facial images and video to classify expressions into categories such as happiness, sadness, anger, fear, surprise, disgust, and neutral. Such systems have been explored for human–computer interaction, audience and event analysis, and other settings in which aggregate visual information may be useful [1]–[6].

Many FER approaches are designed and evaluated on cropped images or individual faces. A crowd scene introduces additional difficulties: multiple faces must be located in each frame, some faces may overlap or be partly hidden, distant faces may contain few pixels, and lighting or head orientation may vary across the scene. A system that classifies a single face accurately does not automatically provide a reliable summary of a whole crowd.

This project proposes an integrated pipeline using YOLO for multi-face detection and a FER component for expression classification. After per-face predictions are produced, the system counts the predicted categories and displays their distribution, along with the dominant predicted category. The project emphasis is integration and crowd-level aggregation rather than proposing a new neural-network architecture.

A. Problem Statement

Traditional or separately evaluated facial-expression systems may not directly address the practical task of detecting several faces and summarising their predicted expressions in a live or recorded crowd video. Detection and classification errors can be introduced by occlusion, overlapping faces, low resolution, changing illumination, unusual head poses, and ambiguous expressions. Processing every frame also creates a speed–accuracy trade-off.

The problem addressed is to design and implement a real-time system that detects multiple faces using YOLO, classifies each detected face using a FER model, and produces an interpretable crowd-level emotion summary while evaluating recognition quality and processing speed.

B. Objectives

The project objectives are:

- accept webcam or pre-recorded video input;

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131114

- detect multiple faces in video frames using YOLO;
- extract face regions and classify expressions using a FER model;
- display face boxes and per-face emotion labels;
- calculate emotion counts, proportions, and the dominant predicted category; and
- evaluate accuracy-related measures and real-time performance under different scene conditions.

The intended project contribution is a practical integration of face detection, expression classification, and crowd-level aggregation in one workflow. The system will also document failure cases and limitations, rather than treating every prediction as ground truth.

II. RELATED WORK

Recent studies have explored deep learning techniques for facial emotion recognition (FER). Alshammari and Alshammari applied YOLOv8 for facial expression detection, while Sareen and Seeja combined YOLOv8 face localization with a pretrained Vision Transformer for seven-class emotion classification [1], [2]. These approaches demonstrate the potential of integrating face detection with emotion classification.

Ma et al. proposed FER-YOLO-Mamba, combining YOLO with a selective state-space model to capture local and long-range facial features [3]. Similarly, EmoNeXt uses an adapted ConvNeXt architecture with spatial transformation, channel attention, and self-attention regularization to improve facial expression recognition [4]. Hu et al. introduced OpenFace 3.0, a lightweight system for facial behaviour analysis, including facial landmarks, action units, and gaze estimation [5]. Devi et al. explored a YOLOv8-based pipeline integrating emotion, age, and gender analysis [6].

Although these studies address face detection and facial analysis, practical crowd emotion monitoring still requires an integrated system that can detect multiple faces, predict their emotions, and summarize the overall emotional distribution. The proposed system combines YOLO-based face detection with FER to analyse webcam streams and recorded videos. Its focus is on integrating individual emotion predictions into a crowd-level summary while considering challenges such as occlusion, low resolution, and changing lighting conditions.

A. Research Gap

The reviewed work provides useful components for face localisation, emotion classification, and broader facial analysis. However, the project brief identifies a practical integration need: a lightweight workflow that detects multiple faces in a scene, attaches an emotion prediction to each detected face, and summarises the distribution across the visible crowd in real time. Many reported results are tied to specific datasets and model configurations, so their numbers cannot be directly compared without a common test protocol.

The gap addressed by this project is therefore an application and integration gap, not a claim that YOLO or FER is new. The project will investigate whether the selected components can operate together on webcam or video input and how crowd-level summaries behave under occlusion, low resolution, changing lighting, and overlapping faces.

B. Proposed Contribution

The proposed system integrates YOLO-based face detection with Facial Emotion Recognition (FER) to detect multiple faces and classify their emotions in real time. It supports analysis of live webcam streams and pre-recorded videos.
The system also generates a crowd-level summary of detected emotions, providing an overall view of emotional distribution. It aims to address practical challenges such as facial occlusion, low resolution, overlapping faces, and varying lighting conditions.

III. SYSTEM OVERVIEW AND ARCHITECTURE

The proposed system integrates YOLO-based face detection with Facial Emotion Recognition (FER) to analyse crowd emotions from live webcam streams and pre-recorded videos. The workflow consists of video input, face detection, emotion classification, and crowd-level emotion analysis. Detected faces are processed to predict emotional expressions, and the results are aggregated to provide an overview of the crowd's emotional distribution.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131114

Layer / Module	Responsibility
Video Input	Captures live webcam streams or loads pre-recorded videos.
Face Detection	Uses YOLO to detect multiple faces in video frames.
Emotion Recognition	Classifies detected facial regions into emotion categories using FER.
Crowd Emotion Analysis	Aggregates individual predictions to summarise the emotional distribution.
Result Visualization	Displays detected faces, emotion labels, and crowd-level results.
Performance Evaluation	Evaluates detection, classification, and processing speed using appropriate metrics.

TABLE I: LAYERS AND RESPONSIBILITIES OF THE PROPOSED SYSTEM

A. Crowd Emotion Detection Workflow

1. The system receives input from a live webcam or a pre-recorded video.
2. Video frames are extracted and passed to the YOLO-based face detection module.
3. Detected facial regions are processed by the FER model to predict emotions.
4. Emotion labels are associated with the corresponding detected faces.
5. Individual predictions are aggregated to generate a crowd-level emotion summary.
6. The results are displayed for observation and analysis.

B. Crowd-Level Emotion Analysis

The system combines individual facial emotion predictions to estimate the overall emotional distribution of the detected crowd. It aims to support analysis under challenging conditions, including overlapping faces, facial occlusion, low image resolution, and variations in lighting. The system's effectiveness should be assessed using suitable detection and classification metrics, along with processing speed under different input conditions.

IV. METHODOLOGY

A. Video Input and Face Detection

The system accepts live webcam streams and pre-recorded videos as input. Video frames are extracted and passed to the YOLO-based face detection module. The model identifies multiple faces within each frame and marks their locations using bounding boxes. The detected facial regions are then passed to the Facial Emotion Recognition (FER) module for further processing. The overall workflow is designed to support continuous video analysis while maintaining a practical balance between detection accuracy and processing speed.

B. Facial Emotion Recognition

The FER module analyses the detected facial regions and classifies them into emotion categories such as happiness, sadness, anger, fear, surprise, and neutrality. Each predicted emotion is associated with its corresponding detected face, allowing multiple individuals to be analysed within a single frame. The effectiveness of emotion classification depends on facial visibility, image quality, lighting conditions, and variations in facial expressions. The model's performance should therefore be assessed under different input conditions rather than assuming consistent accuracy in every environment.

C. Crowd-Level Emotion Analysis

The crowd emotion analysis module combines individual emotion predictions to estimate the overall emotional distribution of the detected crowd. The number or proportion of faces assigned to each emotion category can be used to generate a crowd-level summary. This allows the system to present both individual facial predictions and an aggregated view of the detected emotions. Since facial expressions alone cannot reliably establish a person's actual internal emotional state, the results should be interpreted as predictions of visible facial expressions rather than definitive psychological assessments.

D. Result Visualization

The system presents the processed output by displaying detected faces, their bounding boxes, and corresponding emotion labels. A crowd-level summary can provide an overview of the distribution of predicted emotion categories within the analysed video frames. The visualization component is intended to make the results easier to interpret and observe. The exact presentation of results depends on the implemented interface and the output supported by the detection and emotion recognition modules.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131114

E. Evaluation Plan

The system should be evaluated using appropriate face detection and emotion classification metrics. Detection performance can be measured using precision, recall, and mean Average Precision (mAP), where suitable ground-truth annotations are available. Emotion classification can be assessed using accuracy, precision, recall, and F1-score on an appropriately labelled dataset. Operational evaluation should also consider processing speed, response latency, and performance under challenging conditions such as facial occlusion, overlapping faces, low resolution, and varying illumination. Numerical performance claims should be reported only after completing experiments and documenting the dataset, evaluation procedure, and test conditions.

V. EXPECTED BENEFITS AND DISCUSSION

The proposed system is expected to simplify crowd emotion analysis by integrating YOLO-based face detection and Facial Emotion Recognition (FER) into a unified workflow. Real-time video processing can help identify multiple faces and classify their visible facial expressions, while crowd-level emotion summaries can provide an overview of the predicted emotional distribution. The integration of detection and classification may support applications such as crowd monitoring, audience analysis, and interactive environments. These are expected benefits and should not be considered experimentally established outcomes until the system has been evaluated.

Combining face detection and emotion recognition introduces several technical challenges. Occlusion, overlapping faces, variations in lighting, low image resolution, and different facial orientations may affect detection and classification performance. Incorrect predictions can also influence the overall crowd-level summary. The system should therefore present results as estimated facial expressions rather than definitive measurements of people's internal emotional states. Evaluation under different environmental conditions is necessary to understand the system's reliability and practical limitations.

A. Limitations

The system's performance depends on the accuracy of the YOLO detector, the FER model, and the quality of the input video. Facial occlusion, crowded scenes, and poor illumination may reduce prediction reliability. Real-time processing performance may also vary depending on hardware capabilities and video resolution. Furthermore, emotion categories predicted from facial expressions may not accurately represent an individual's actual emotional state. The system should therefore be treated as an emotion analysis aid rather than a definitive psychological assessment tool. Performance claims must be supported by experimental evaluation.

VI. FUTURE SCOPE

- Implement document ingestion, chunking, embeddings, vector search, and source-aware retrieval for both interview and quiz workflows.
- Develop quiz generation with configurable question count, topic, difficulty, question format, answer key, and explanation.
- Add automated checks for answer-key consistency, duplicate questions, source support, and distractor validity, followed by human review.
- Store learner progress securely and optionally use previous practice topics to recommend revision areas.
- Evaluate the application with students using predefined rubrics and report relevance, correctness, usability, latency, and failure cases.
- Explore multilingual content and role-specific knowledge bases while assessing the impact of domain coverage and retrieval quality.

VII. CONCLUSION

This paper presents a system design for a Real-Time Crowd Emotion Detector Using YOLO and Facial Emotion Recognition (FER). The proposed approach combines face detection, facial emotion classification, and crowd-level emotion analysis to process live webcam streams and pre-recorded videos. The system aims to identify multiple faces, predict their visible facial expressions, and summarize the overall distribution of predicted emotions.

The main objective is to integrate face detection and emotion recognition into a unified workflow for crowd analysis. Further implementation and systematic evaluation are required to assess its accuracy, processing speed, and reliability under different environmental conditions. The system's predictions should be interpreted as estimates of visible facial expressions rather than definitive measurements of individuals' emotional states.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

DOI: 10.17148/IARJSET.2026.131114

REFERENCES

[1]. A. Alshammari and M. E. Alshammari, "Emotional Facial Expression Detection using YOLOv8," Engineering, Technology & Applied Science Research, vol. 14, no. 5, pp. 16619–16623, 2024, doi: 10.48084/etasr.8433.

[2]. V. Sareen and S. K. R. Seeja, "Video-Based Facial Emotion Recognition using YOLO and Vision Transformer," in the supplied merged research-paper collection. Full publication details require verification.

[3]. H. Ma, S. Lei, T. Celik, and H.-C. Li, "FER-YOLO-Mamba: Facial Expression Detection and Classification Based on Selective State Space," arXiv preprint arXiv:2405.01828, 2024.

[4]. Y. El Boudouri and A. Bohi, "EmoNeXt: An Adapted ConvNeXt for Facial Emotion Recognition," arXiv preprint arXiv:2501.08199, 2025.

[5]. J. Hu, L. Mathur, P. P. Liang, and L.-P. Morency, "OpenFace 3.0: A Lightweight Multitask System for Comprehensive Facial Behavior Analysis," arXiv preprint arXiv:2506.02891, 2025.

[6]. Y. Ming, H. Qian, and G. Liu, "CNN-LSTM Facial Expression Recognition Method Fused with Two-Layer Attention Mechanism," Computational Intelligence and Neuroscience, vol. 2022, Art. no. 7450637, 2022, doi: 10.1155/2022/7450637.

[7]. R. Singh et al., "Facial Expression Recognition in Videos Using Hybrid CNN & ConvLSTM," International Journal of Information Technology, vol. 15, no. 4, pp. 1819–1830, 2023, doi: 10.1007/s41870-023-01183-0.

[8]. "Facial Expression Recognition Using Bidirectional LSTM-CNN," Procedia Computer Science, vol. 216, pp. 39–47, 2023, doi: 10.1016/j.procs.2022.12.109. Author details should be checked against the original paper.

[9]. "Real-Time Facial Expression Recognition to Enhance Emotional Intelligence in Autism," Procedia Computer Science, vol. 234, pp. 222–229, 2024, doi: 10.1016/j.procs.2024.02.169. Author details should be checked against the original paper.

[10]. B. H. Pansambal, A. B. Nandgaokar, J. L. Rajput, and A. Wagh, "An Integrated CNN-BiLSTM Approach for Facial Expressions," International Journal of Advanced Computer Science and Applications, vol. 15, no. 3, 2024, doi: 10.14569/IJACSA.2024.0150398.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License