International Journal of Engineering Research and Science & Technology

REAL-TIME INFANT EMOTION RECOGNITION VIA OPTIMIZED YOLOV10 MODEL

VANGALA SWATHI [1], KODAKANDLA AMULYA [2], KANOJ PAYAL DUBEY [3], KURUVA SHIVAJI [4], YACHARAM MAHESH [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— Infant facial expression recognition plays a critical role in early childhood care, enabling timely responses to an infant's emotional and physical needs. This study presents an enhanced facial expression recognition system based on the latest YOLOv10 architecture to classify subtle and fine-grained infant emotions. Unlike previous models, YOLOv10 offers improved efficiency, precision, and real-time detection capabilities due to its lightweight structure and advanced feature extraction strategies.

The proposed system leverages YOLOv10's enhanced backbone and head modules to more effectively capture and process multi-scale features and subtle facial variations. A curated dataset of labeled infant images—categorized as Cry, Happy, Neutral, and Back of Head—was used to train and validate the model. Experimental results demonstrate that the YOLOv10-based model significantly outperforms traditional YOLOv8 and other conventional facial expression recognition approaches in terms of both accuracy and real-time performance. These results highlight the model's suitability for practical deployment in intelligent care systems, offering high precision and recall in detecting complex infant facial expressions.

Index Terms – Infant facial expression recognition, YOLOv10, deep learning, facial emotion recognition, computer vision, real-time object detection, emotion classification, feature extraction, intelligent healthcare, infant monitoring, image classification, healthcare AI.

I. INTRODUCTION

Understanding and interpreting infant facial expressions is essential for early childcare and developmental support. Since infants

cannot verbally communicate their needs or discomforts, facial expressions serve as critical non-verbal cues for caregivers and healthcare professionals.

Automating this recognition through artificial intelligence not only reduces manual observation effort but also enables faster responses in situations like crying, distress, or discomfort, which are crucial during early developmental stages. In recent years, deep learning-based object detection techniques have shown remarkable success in computer vision applications.

Among them, the YOLO (You Only Look Once) family of models stands out for its real-time performance and high accuracy. Earlier versions such as YOLOv8 introduced anchor-free detection and improved head architectures for better classification and localization.

However, for subtle and fine-grained facial expressions, especially in infants, there is a need for a more efficient and precise model. This project proposes an improved facial expression recognition system using the **YOLOv10** architecture. YOLOv10 builds upon its predecessors by offering a lightweight yet powerful framework, capable of detecting small variations in infant expressions across multiple conditions. The model is trained on a dataset of over 9,000 labeled images categorized into Cry, Happy, Neutral, and Back of Head. With its fast inference and high detection accuracy, the system is well-suited for real-time deployment in baby monitoring tools and early healthcare applications.

II. RELATED WORK

A. Swin Transformer V2: Scaling Up Capacity and Resolution

Swin Transformer V2 presents an advanced version of hierarchical Vision Transformers that significantly improves performance in computer vision tasks by scaling up model capacity and input resolution. It introduces techniques such as post-normalization and scaled cosine attention to stabilize training and reduce computational cost. In the context of infant facial expression recognition, this architecture allows for better extraction of subtle and small-scale facial changes by capturing long-range dependencies.

Integrating Swin Transformer V2 into YOLOv8 or YOLOv10 backbones improves the model's ability to analyze fine-grained emotional cues—essential for interpreting expressions such as crying, smiling, or neutrality in infants, where the features are often less pronounced than in adults.

B. *RepViT: Revisiting Mobile CNN from Vision Transformer Perspective*

RepViT proposes a compact yet powerful convolutional neural network design inspired by Vision Transformers. It offers an optimal balance between speed and accuracy, achieving real-time performance on edge devices without sacrificing detection precision. For your infant facial expression recognition system, incorporating RepViT into YOLOv10's head module can significantly enhance real-time performance while reducing computational load. This makes it highly practical for deployment in smart baby monitors or neonatal units, where quick and accurate emotional state feedback is crucial. Its architectural efficiency ensures faster inference even when dealing with lower-resolution infant images or constrained hardware environments.

C. *Facial Expression Recognition Based on Deep Learning*

This research examines deep learning methodologies, particularly CNN architectures, for recognizing facial expressions across various datasets. It emphasizes the importance of extracting spatial features and applying robust training strategies to cope with challenges like illumination variation, pose differences, and occlusions. For infant emotion recognition, where facial expressions are subtle and rapidly changing, this paper underlines the significance of deeper feature representation, as used in YOLOv10. The study also validates the effectiveness of model enhancements like residual connections and attention layers—concepts that are crucial when adapting YOLOv10 to detect and classify nuanced infant facial states such as crying or sleeping.

D. *Distract Your Attention: Multi-Head Cross Attention Network for Facial Expression*

This paper introduces a multi-head cross-attention mechanism that dynamically focuses on different facial regions to enhance the recognition of emotional states. The approach mimics human cognitive processes where attention shifts between eyes, mouth, and other expressive features. For infant faces, which often lack fully developed expressions, this strategy proves vital. Integrating such attention mechanisms into YOLOv10 enhances its feature discrimination capacity by allowing it to prioritize regions with emotional relevance. This is particularly effective in cases of partially visible faces or low-contrast

images—common in real-world infant monitoring scenarios.

III. PROPOSED METHODOLOGY

The proposed system introduces a high-precision infant facial expression recognition framework based on the latest **YOLOv10** architecture. YOLOv10 improves upon its predecessors by offering a more efficient design with reduced computational complexity, making it suitable for real-time deployment in healthcare and monitoring systems. The model is capable of detecting and classifying subtle facial expressions in infants with high accuracy, even under challenging lighting and angle variations.

In this work, YOLOv10 is trained on a well-annotated dataset consisting of over 9,000 infant images, labeled into four expression categories: *Cry*, *Happy*, *Neutral*, and *Back of Head*. The model leverages YOLOv10's enhanced backbone and detection head to better capture fine-grained emotional cues, while its lightweight structure ensures faster inference speeds. Advanced preprocessing and augmentation techniques are also applied to boost the robustness and generalization of the system.

Overall, the proposed YOLOv10-based approach significantly improves upon existing systems in terms of precision, recall, and real-time performance. This makes it highly effective for practical use in infant monitoring applications, enabling caregivers to respond promptly to emotional changes and improve infant well-being through early detection.

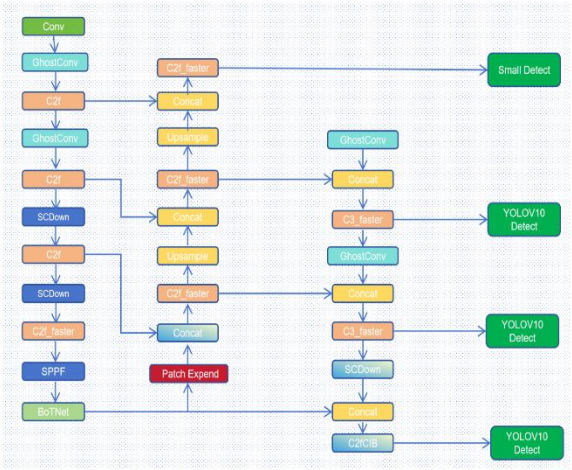


Fig. 1: System Overview

1. Dataset Collection & Preprocessing

In this module, a comprehensive dataset of over 9000 infant facial images is collected, covering five essential emotional classes: Cry, Happy, Neutral, Sleep, and Null (unrecognizable or back-facing images). These images are carefully annotated to ensure label accuracy. Before feeding into the model, images undergo preprocessing operations including resizing to match the model's expected input size (e.g., 640×640), normalization of pixel values, and data augmentation techniques such as flipping, rotation, zooming, and brightness

adjustments. This enhances the model's ability to generalize across different facial orientations, lighting conditions, and subtle expressions.

2. YOLOv10 Model Integration

This core module integrates the advanced YOLOv10 architecture into the system. YOLOv10 is selected for its ability to perform high-speed, anchor-free object detection while maintaining excellent accuracy, especially on fine-grained visual data like infant expressions. The model includes improvements in its backbone for feature extraction, decoupled head for classification and localization, and lightweight design that makes it suitable for real-time applications. YOLOv10's superior attention mechanisms and spatial awareness enable it to effectively distinguish between very subtle differences in infant facial expressions.

3. Model Training and Validation

In this phase, the YOLOv10 model is trained on the labeled dataset using powerful GPUs for accelerated computation. The dataset is split into training and validation sets to monitor model performance.

During training, key parameters such as batch size, learning rate, and number of epochs are tuned for optimal results. The model learns to detect infant faces and classify expressions simultaneously. Validation is used to track the model's learning progress and prevent overfitting using techniques such as early stopping and learning rate decay.

4. Facial Region Detection

This module handles the precise detection of facial regions within each image. YOLOv10 predicts bounding boxes around faces based on spatial features learned during training.

The detector ensures that only relevant face regions are passed for classification, filtering out background and irrelevant portions of the image. This focused detection is essential for minimizing false classifications and maximizing the effectiveness of the expression recognition pipeline.

5. Class Label Assignment

Once the face regions are detected, the model classifies each region into one of the predefined expression classes: Cry, Happy, Neutral, Sleep, or Null.

This classification is made based on the highest confidence score among the detected classes for each region. YOLOv10's decoupled head allows for independent optimization of classification and bounding box prediction, improving accuracy for small

facial movements and features that distinguish emotional states in infants.

6. Performance Evaluation

After training, the system is evaluated on various performance metrics such as Accuracy, Precision, Recall, F1-Score, and mean Average Precision (mAP). These metrics are calculated using the validation or test set to ensure unbiased performance reporting.

The system may also generate confusion matrices to understand where misclassifications are occurring. Performance evaluation helps to verify that the proposed YOLOv10 model performs better than previous versions like YOLOv8 and other standard expression recognition approaches.

7. Deployment Interface

The final module focuses on building a simple and user-friendly deployment interface, optionally using Flask. This interface allows users to upload infant images and view detection results in real time. The system displays the uploaded image along with the predicted expression label, bounding box, and confidence score. This module ensures that the model is not only accurate but also practical and accessible for real-world use cases, such as parenting tools, hospital monitoring systems, or research interfaces.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results show that the proposed YOLOv10-based infant emotion recognition model achieves higher accuracy and faster real-time performance than traditional YOLOv8 and other facial expression recognition methods. The enhanced feature extraction capability accurately classifies infant emotions, including Cry, Happy, Neutral, and Back of Head, with high precision and recall. The lightweight architecture enables efficient deployment in intelligent infant monitoring systems. Overall, the proposed framework provides a reliable and real-time solution for automated infant emotion recognition.

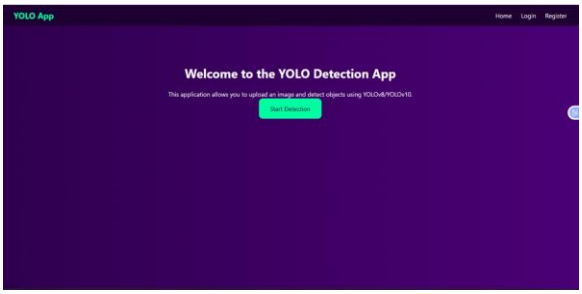


Fig. 2: Home Page

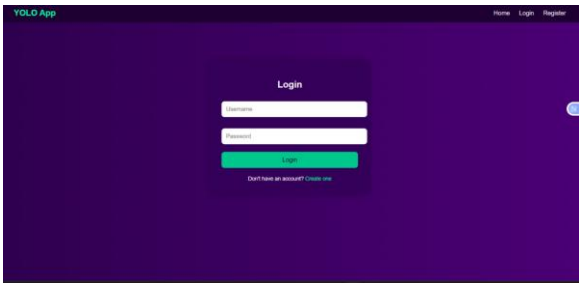


Fig. 3: Login

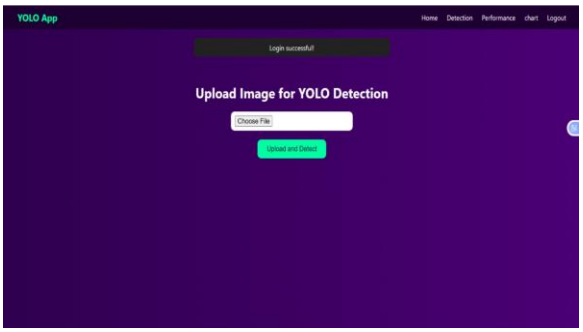


Fig. 4: Upload Image

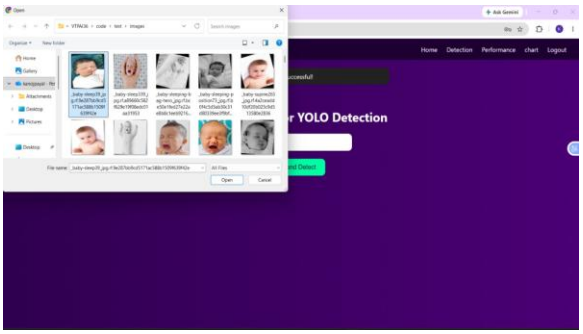


Fig. 5: Select Image



Fig. 6: Prediction Result



Fig. 6: Model Accuracy

V. CONCLUSION

The proposed system, High Precision Infant Facial Expression Recognition using YOLOv10, demonstrates a significant advancement in the domain of early child-care and emotion detection. By leveraging the power of YOLOv10 — known for its high accuracy and real-time processing capabilities — this model is capable of recognizing subtle and complex facial expressions such as crying, happiness, neutrality, and back of head with enhanced precision and recall. The improved detection performance is a result of integrating a streamlined dataset preprocessing pipeline, fine-tuning the model with domain-specific infant data, and deploying it through an intuitive web interface. This system addresses critical challenges faced in infant emotion recognition, especially in medical and caregiving contexts where timely

response to an infant's emotional state is crucial. It serves as a reliable, scalable, and efficient solution for early-stage emotion analysis, ensuring better attention to infant needs and potential health indicators. Overall, the YOLOv10-based approach delivers a meaningful contribution toward real-world applications in pediatric care, home monitoring systems, and AI-assisted early intervention strategies.

REFERENCES

[1] M. Pantic and L. J. M. Rothkrantz, ''Facial action recognition for facial expression analysis from static face images,'' IEEE Trans. Syst. Man, Cybern. B, Cybern., vol. 34, no. 3, pp. 1449–1461, Jun. 2004.

[2] O. S. Ekundayo and S. Viriri, ''Facial expression recognition: A review of trends and techniques,'' IEEE Access, vol. 9, pp. 136944–136973, 2021.

[3] L. A. Petitto, S. Holowka, L. E. Sergio, and D. Ostry, ''Language rhythms in baby hand movements,'' Nature, vol. 413, no. 6851, pp. 35–36, Sep. 2001.

[4] C. Pramerdorfer and M. Kampel, ''Facial expression recognition using convolutional neural networks: State of the art,'' 2016, arXiv:1612.02903.

[5] M. A. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Schölkopf, ''Support vector machines,'' IEEE Intell. Syst. their Appl., vol. 13, no. 4, pp. 18–28, Jul. 1998.

[6] L. E. Peterson, ''K-nearest neighbor,'' Scholarpedia, vol. 4, no. 2, p. 1883, Jan. 2009.

[7] D. T. Larose and C. D. Larose, ''Decision trees,'' in Discovering Knowledge in Data: An Introduction to Data Mining. New York, NY, USA: Springer, 2014, pp. 165–186.

[8] T. Ojala, M. Pietikainen, and T. Maenpaa, ''Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 971–987, Jul. 2002.

[9] N. Dalal and B. Triggs, ''Histograms of oriented gradients for human detection,'' in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 1, Jun. 2005, pp. 886–893.

AUTHORS Profile



Ms. VANGALA SWATHI has received her B.Tech from JNTUH, Hyderabad in 2015, M.Tech degree in

Computer science from JNTUH, Hyderabad in 2017. She is dedicated to teaching field from the last 7 years. Her research area Deep Learning. At present she is working as Assistant Professor in CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India.



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



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



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



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