Emotion Mapping through Facial Markers: Harnessing CNNs and RNNs for Real-Time Recognition.

Sudhanshu Lacca1, Piyush Nikam2, Mayur Joshi3, Aditya Walke4, Komal M. Birare5

*Department of Information Technology, Vishwakarma Institute of Information Technology, Pune, Maharashtra, India*

[sudhanshu.lacca@gmail.com](mailto:1sudhanshu.22211574@viit.ac.in)1, [nikampiyush2003@gmail.com](mailto:nikampiyush2003@gmail.com)2, [mayurjoshi1309@gmail.com](mailto:mayurjoshi1309@gmail.com)3, [walkleaditya3@gmail.com](mailto:walkleaditya3@gmaail.com)4, [komal.birare@vit.edu](mailto:komal.birare@vit.edu)5

***Abstract:* This paper presents a novel real-time detection system, combining pre-trained models of state-of-the-art deep learning for emotion recognition, age and gender prediction, and object detection. The design of the age and gender-prediction system utilizes pre-trained deep learning models. Emotion detection utilizes the state-of-the-art CNN model. The system employs the state-of-the-art YOLOv8 object detector to obtain and display multiple objects in live video feeds effectively and accurately. The application uses OpenCV for the in-built webcam stream, utilizing a Streamlit-based user interface so it is interactive. Some major features include options to turn grayscale on/off, object detection enabled/disabled, and predictions rendered in real-time, all with optimized performance by calculating FPS and using modular components that can be easily scaled up. Applications of this work include human-computer interaction, security systems and analysis of behavior. The proposed system has multi-detection capabilities channeled into a single package; and hence, it is a tool that is all-inclusive of multi-facet real-time analysis. The system was demonstrated to be effective in live environments, and further improvements could be achieved using acceleration through the GPU and fine-tuning the model.**

***Keywords*: Real-time detection, emotion recognition, age and gender prediction, object detection, YOLOv8, CNN, OpenCV, Streamlit, behavioral analysis, human-computer interaction**.

1. INTRODUCTION

Within the last few years, computer vision and deep learning have been developed, making it possible to establish intelligent systems that can perform highly complicated tasks in real time, such as emotion detection, age and gender estimation, and object detection. It has applied the technologies in different areas, including security, human-computer interaction, retail analytics, and behavioral studies. It introduces the unified real-time detecting system, from which various capabilities are brought together into one single yet efficient platform. The system also utilizes the CNN for emotion detection, pre-trained deep models for age and gender prediction, and finally the YOLOv8 model, for accurate real-time object detection in live video feeds.

The use of OpenCV is for webcam integration, allowing real-time video feed processing; the use of Streamlit is employed to provide an intuitive and interactive user interface. The customization options here are with modes such as grayscaling to enhance face detection and toggle-able mode for object detection; in addition, performance monitoring has been conducted in real-time by calculating FPS. Artificial intelligence and deep learning marked a new era in the landscape of computer vision and have been used to facilitate real-time data extraction towards meaningful insights from visual data. In this domain, impactful tasks range from emotion recognition, age and gender prediction, object detection across security systems, retail analytics, healthcare, smart environments, and human-computer interaction. Despite such giant progress, such integration is still highly challenging due to computational constraints and a big requirement for high accuracy and complexity in the combining of heterogeneous models. [5].

The paper discusses an integrated system implemented by state-of-the-art deep learning models that recognize emotions, predict age and gender, as well as detect objects with real-time processing in the single framework. It runs emotion detection on a convolutional neural network powered by facial expressions, the age and gender model pre-trained for demographic insights and uses the YOLOv8 object detection model to deliver speeds that cannot be matched and probably accuracy in terms of detecting objects in feeds. All these models combined, put together provide a robust multi-tasking system fit to work well in dynamic environments. This would entail streaming video, processing it in real time using OpenCV, and a Streamlined interface to engage users. This system has controls over toggleable features like object detection and grayscale mode, but to cater to all user needs, there are controls for usability and efficiency use-through performance metrics such as FPS and the number of detected faces. This system, which integrates numerous functionalities in a single platform, addresses the demands for diversified real-time solutions in such applications as behavioral analysis, surveillance, and interactive kiosks. The modularity also makes it scalable, which lays ground for future advances such as accelerator deployment based on GPU acceleration, updating the model, among others. This paper presents the architecture, implementation of this system, and its possible applications, thus demonstrating its significance in today's world of data.[8]

Objectives :

1. Emotion Detection

Emotion detection involves analyzing facial expressions from live video feeds to identify emotions such as happiness, sadness, anger, surprise, and more. Using advanced computer vision and deep learning techniques, the system decodes facial cues to provide real-time insights into human emotional states. This feature is valuable for understanding user behavior and improving interaction experiences.

1. Age and Gender Detection

Age and gender detection estimates a person’s age range and identifies their gender based on facial features. Deep learning models trained on diverse datasets enable accurate predictions, even in dynamic environments. This capability is widely applicable in areas like targeted marketing, security, and audience analysis.

1. Object Detection

Object detection identifies and localizes objects within a video frame or image. The system uses machine learning algorithms to classify and highlight objects such as furniture, vehicles, or accessories. This feature is useful in applications ranging from inventory management to automated surveillance.

1. Detection of Confidence Score

The system provides a confidence score for every prediction it makes, indicating the model's certainty about its results. This score helps users assess the reliability of the detected emotion, age, gender, or object. Higher confidence levels indicate more accurate predictions, aiding in decision-making processes.

1. Integration with Streamlit Web App

The system integrates seamlessly with a Streamlit-based web application for a user-friendly interface. Streamlit allows real-time visualization of the system’s outputs, such as detected emotions, demographics, objects, and confidence scores. This integration ensures easy accessibility and interactivity, making the tool versatile and intuitive for end users.[11].

II. LITERATURE SURVEY

This paper explores both conventional machine learning methods (e.g., SVM, HMM, KNN) and advanced deep learning techniques (e.g., CNN, CNN-RNN) for facial emotion recognition. Key metrics such as precision, recall, and F1-score are used to evaluate the models' performance.

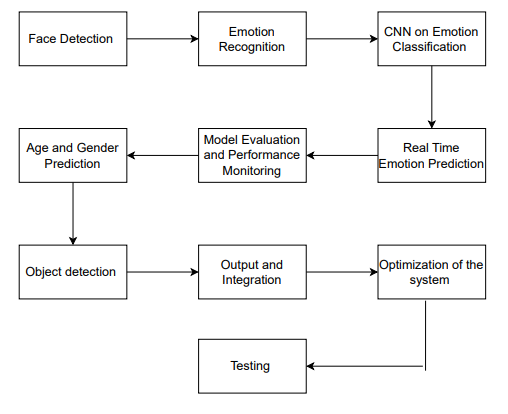
The study highlights the limitations of existing systems in capturing nuanced emotions beyond the basic six categories.[1]. It also emphasizes the challenge of acquiring large labeled datasets for training accurate models.

Future work could focus on incorporating more complex emotional states into FER systems and developing methods to handle the data scarcity problem efficiently.

| **Paper** | **Techniques(s) used** | **Key Points in Coverage** | **Areas for Further Investigation** |
| --- | --- | --- | --- |
| Facial Emotion Recognition Using Convolutional Neural Networks | The paper proposes a deep learning model using Convolutional Neural Networks (CNN) for facial emotion recognition. The model consists of six convolutional layers, two max-pooling layers, and two fully connected layers. | Accuracy: The CNN model achieved an accuracy of 60% after tuning various hyperparameters, such as learning rate and number of filters.  Comparative Performance: CNN outperformed other models like decision trees and fully connected neural networks in terms of accuracy for image recognition tasks. | Future work could focus on further tuning of hyperparameters or experimenting with different network architectures to increase accuracy.  The implementation of the model in real-time applications, such as human-computer interaction systems, could be explored. |
| Facial emotion recognition using convolutional neural networks (FERC) | FERC uses a two-part CNN where the first part removes the background, and the second part extracts facial features. The extracted expressional vectors (EV) are used for classifying emotions. | The model achieved a 96% accuracy in detecting emotions, outperforming single-level CNN approaches. | Extending Emotion Categories: Research can be conducted to include a broader range of emotions beyond the five basic ones used in this study.  Application in Real-World Systems: Real-world application scenarios, such as driver monitoring systems, can be explored with FERC for emotion detection |
| Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges | Conventional machine learning approaches (e.g., using classifiers like SVM, HMM, KNN, and RF on extracted facial features) and deep learning-based approaches (e.g., employing CNNs and hybrid architectures like CNN-RNN or CNN-LSTM). | The sources discuss the performance of FER systems using various metrics such as precision, recall, F1-score, true negatives (TN), false negatives (FN), true positives (TP), and false positives (FP). | The sources acknowledge that most current FER systems focus on recognizing a limited set of basic emotions (e.g., happiness, sadness, anger, fear, surprise, disgust). However, human emotional experience is far more nuanced and complex. Training accurate deep learning models typically requires large amounts of labeled data, which can be expensive and time-consuming to acquire. Future efforts could explore |
| A study on computer vision for facial emotion recognition | This study utilizes a deep neural network (DNN) combining a squeeze-and-excitation network (SENet) and a residual neural network (ResNet) for facial emotion recognition (FER).[3] | The study reports achieving a 77.37% accuracy when the model trained on AffectNet is used to predict emotions in the RAF-DB dataset and a validation accuracy of 83.37% when transfer learning is applied to the RAF-DB with a model pre-trained on AffectNet. The study also notes that the accuracy on AffectNet is 56.54% and the accuracy on RAF-DB without pre-training is 65.67%. | Future research could focus on developing FER systems capable of recognizing and interpreting a broader range of human emotions, encompassing more subtle emotions and complex emotional states. |
| Facial Emotion Recognition and Classification Using the Convolutional Neural Network-10 (CNN-10) | CNN-10, Vision Transformer (ViT), InceptionV3, and VGG19 models. The study compares the performance of these models on three datasets | The CNN-10 model achieved 99.9% accuracy on the CK+ dataset, 84.3% accuracy on FER-2013, and 95.4% accuracy on the JAFFE dataset for facial emotion recognition.[2] | The sources point out that CNNs are often considered "black-box" models, making it challenging to interpret their decision-making process. Future research could explore methods to make facial emotion recognition systems more transparent and interpretable, such as incorporating attention mechanisms. |
| Emotion Recognition using Deep Learning Techniques’ | The paper describes the detection of emotions with the help of different deep learning approaches with an emphasis on CNN or the Convolutional Neural Network. | All the models provided high accuracy rates when tested on various datasets, and these outperformed other conventional methods. | Analyzing performance of training models regarding various datasets in order to increase impact of training. |

III. Methodology

This methodology integrates multiple deep learning models to perform real-time emotion recognition, age and gender prediction, and object detection using a modular system. Face detection is achieved with Haar Cascade Classifier, while emotion recognition relies on a CNN trained on FER2013, and object detection uses YOLOv8. The system optimizes performance with techniques like class imbalance handling, parallel processing, and GPU acceleration, ensuring efficient and robust multi-task execution on live video streams.[4]



1. **Face Detection**

The system employs OpenCV's implementation of Haar Cascade Classifier to detect the faces in video frames. This method detects regions corresponding to faces and crops them for further processing. For the module for emotion recognition, the cropped faces were converted to grayscale, re-sized to 48 × 48 pixels, while for the age and gender prediction models, they are resized to 227 × 227 pixels to meet the necessary conditions in the respective models.

1. **Emotion Recognition**

A CNN model is trained on an emotion dataset for the purposes of facial expression classification in terms of happy, sad, angry, and neutral. The model runs on grayscale face images, giving them probabilities for every emotion. Over the video feed, the highest-probability emotion is overlaid in real-time for user feedback.

1. **Convolutional Neural Networks on Emotion Classification**

The next step includes emotion classification, which is carried out using a pre-trained CNN model passed through a pre-processed image. The FER2013 dataset was involved for the training of this model that is programmed to be implemented for detecting seven types of emotions categorized into: angry, disgust, fear, happy, neutral, sad, and surprise. The CNN architecture consists of several convolutional layers, which capture features such as edges and patterns from the input image, and a number of max-pooling layers, that compress the spatial dimensions of the data to make the system computationally more efficient. Finally, the obtained features are forwarded to fully connected layers that perform the classification task based on these mappings to the respective emotion categories. In the output layer of CNN, we have used a softmax activation function to give the probability of each emotion class, and it has chosen the class with the highest probability as the predicted emotion. The system loads this pre-trained model by loading the architecture model from a JSON file and then it loads the weights from the H5 file.[7].

1. **Real-Time Emotion Prediction**

The last stage is real-time emotion prediction. For every frame captured by the webcam, the system first detects the face and preprocesses the image to send to the CNN for emotion prediction. After making a prediction, the model, the system overloads the predicted label onto the video feed near the bounding box that encloses the detected face. The real-time prediction is of prime importance to the system which can interactively identify and transmit emotions among people inside the video stream.

1. **Model Evaluation and Performance Monitoring**

Continuous testing of the emotion recognition system will prove helpful in assuring the model to work perfectly. The performance of the models can be tested by diverse video inputs at various angles, different lighting conditions, and on facial expressions. Accuracy, precision, recall, and F1-score are measured in terms of the metrics, which show measurable quantities about the performance of the model. If it classifies a few wrongly, the patterns acquired from such analysis would further allow improvements either in the stage of model training or in the preprocessing stages.

1. **Age and Gender Prediction**

It was predicting the age and gender of the detected faces based on pre-trained models; for example, based on Caffe framework. Gender prediction was considered as a binary classification task - Male/Female; while age prediction was done by categorizing persons into predefined age groups. The cropped face images were converted into blobs to feed into models in order to be able to infer. Results of predictions were shown alongside the results of emotion recognition.

1. **Object Detection**

A highly advanced object detection model, YOLOv8 is used for the identification and classification of objects within the video stream. Object detection is done on the entire video frame by drawing bounding boxes over the objects and labeling them with class names and confidence scores. On deployment, multiple object detections call upon real-time capabilities and accuracy in YOLOv8.[10].

1. **Output and Integration**

The combined results from emotion recognition, age and gender prediction, and object detection are presented in real time on the video feed. Detected faces are labeled with emotions, age, and gender, and detected objects are labeled with bounding boxes and their class names. This modularity lends itself well to efficient execution of multiple tasks on live video streams.

1. **Optimization of the System**

Class Imbalance Handling: The approach includes weighted cross-entropy loss in emotion recognition tasks when some emotions might not have occurred often in the dataset.

Parallel Processing: It makes extensive use of multi-threading for optimizing video frame processing and thus for real-time performance.

GPU: It uses the available GPU if possible to speed up the computation, especially for deep learning inference operations.

1. **Testing**

The system performs using tests such as accuracy, precision, recall, and F1-score on each task. For the case of object detection, metrics such as the mean Average Precision (mAP) are considered. Testing is performed on separate datasets in order to evaluate the generalization capabilities of the models and to ensure their robustness in a real-world scenario.[6].

IV. RESULTS AND DISCUSSION

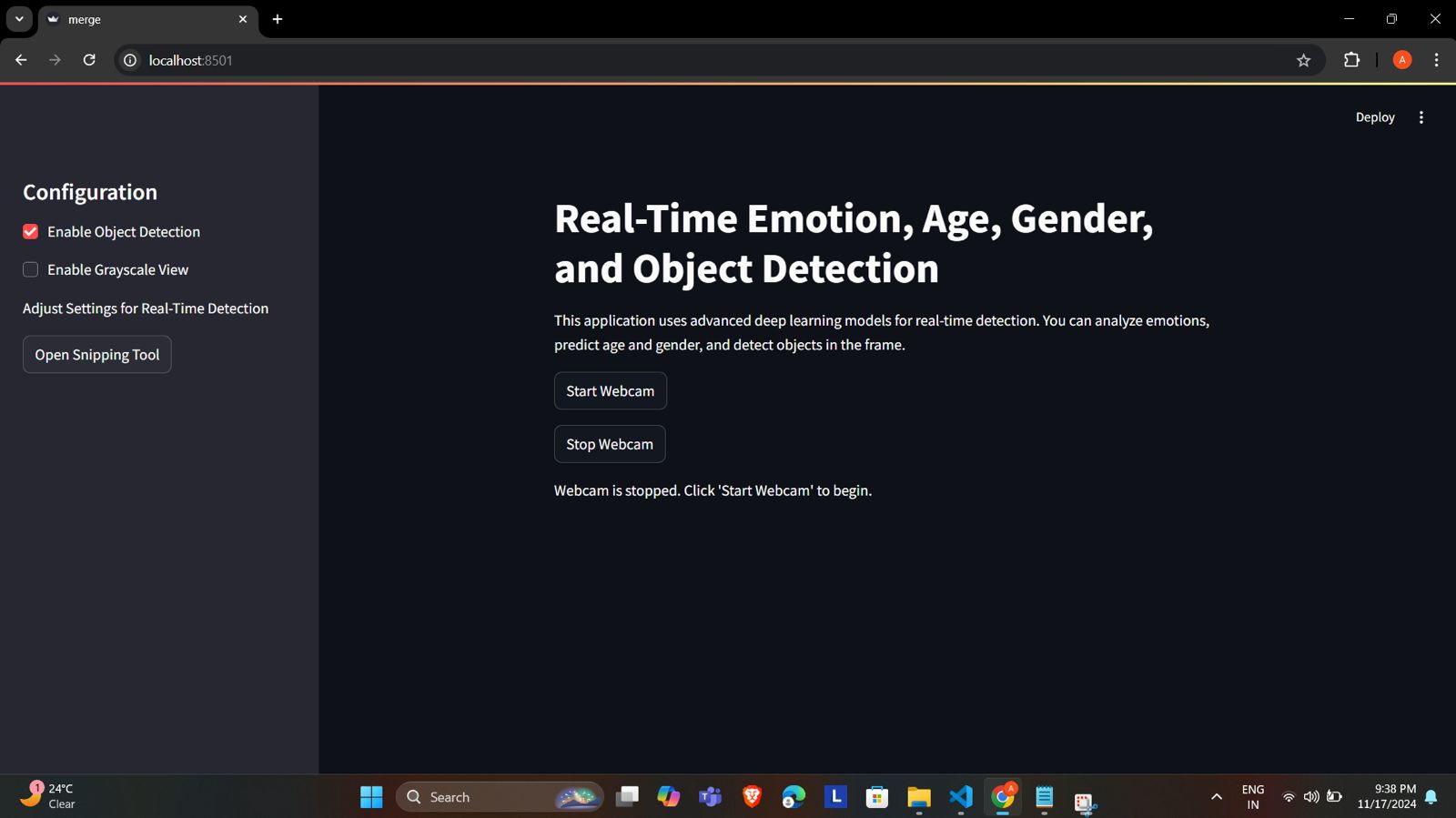


Fig. 1 Web application interface for Real-Time Emotion, Age, Gender, and Object Detection

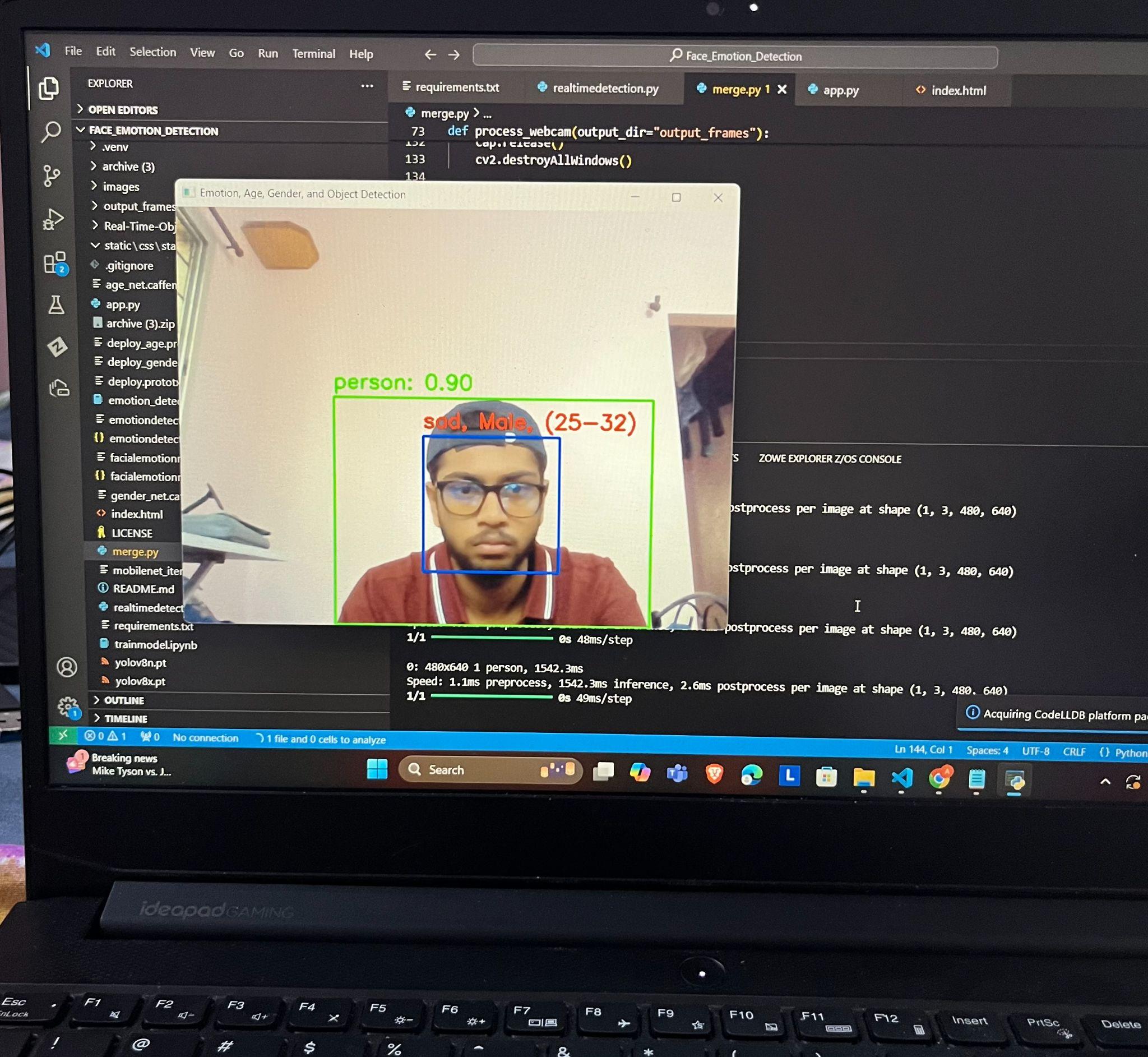


Fig. 2 The image shows a laptop screen running a Python-based real-time detection application where emotion, gender and age have been detected.

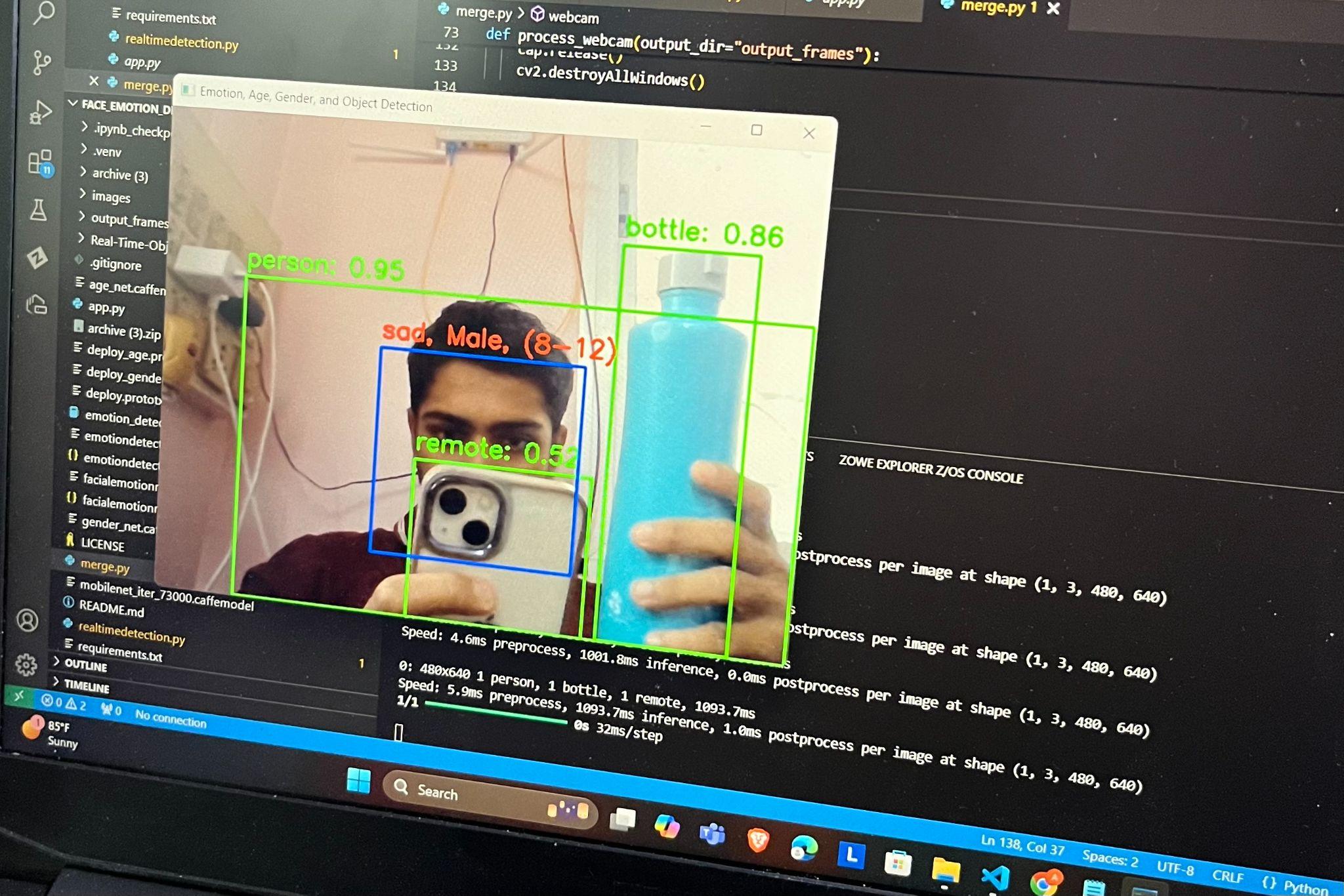


Fig. 3 The image shows a laptop screen running a Python-based real-time detection application where along with emotion, gender and age, objects have also been detected.

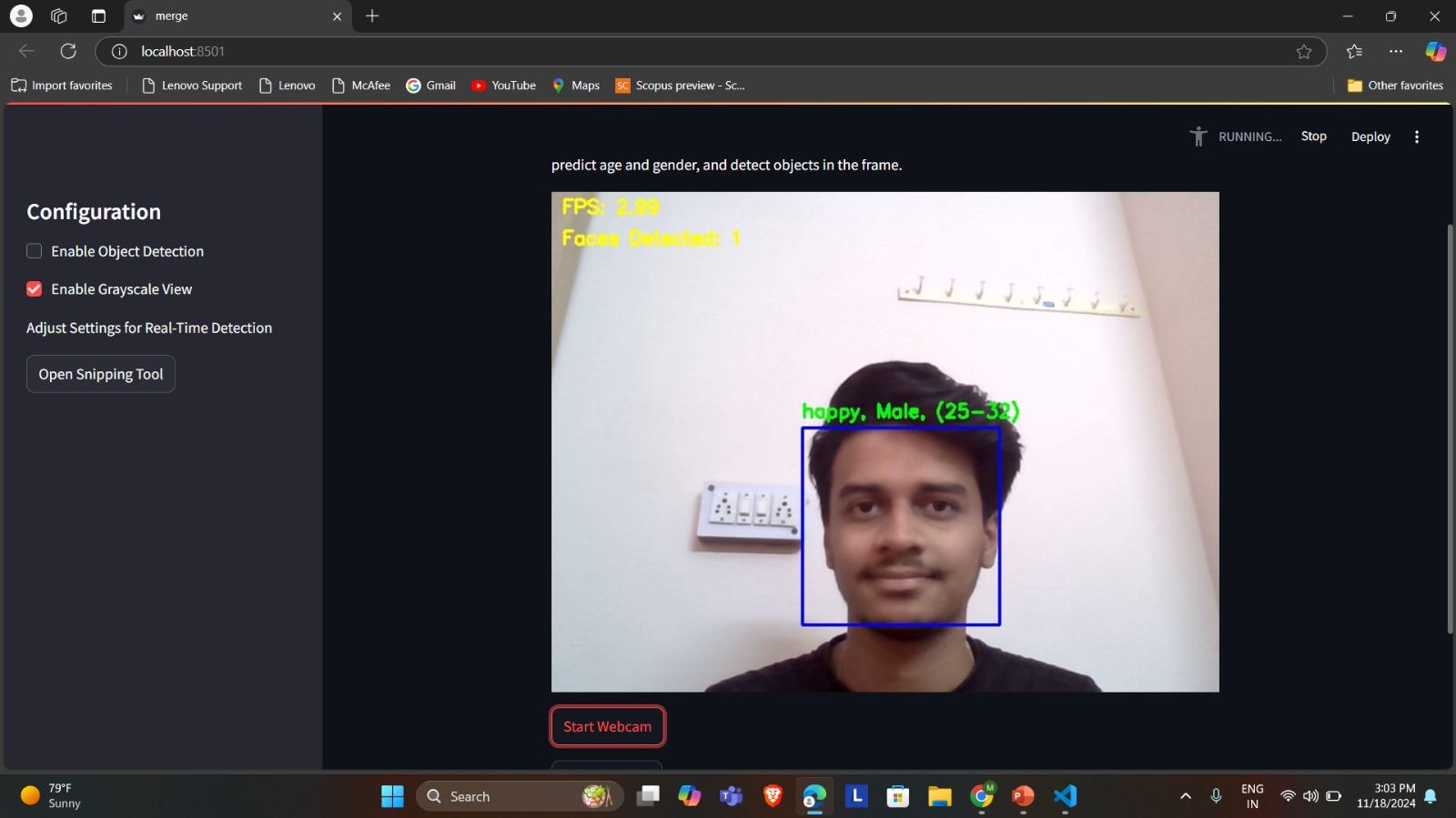


Fig. 4 Grayscale view which enables to only detect the emotions, age and gender.

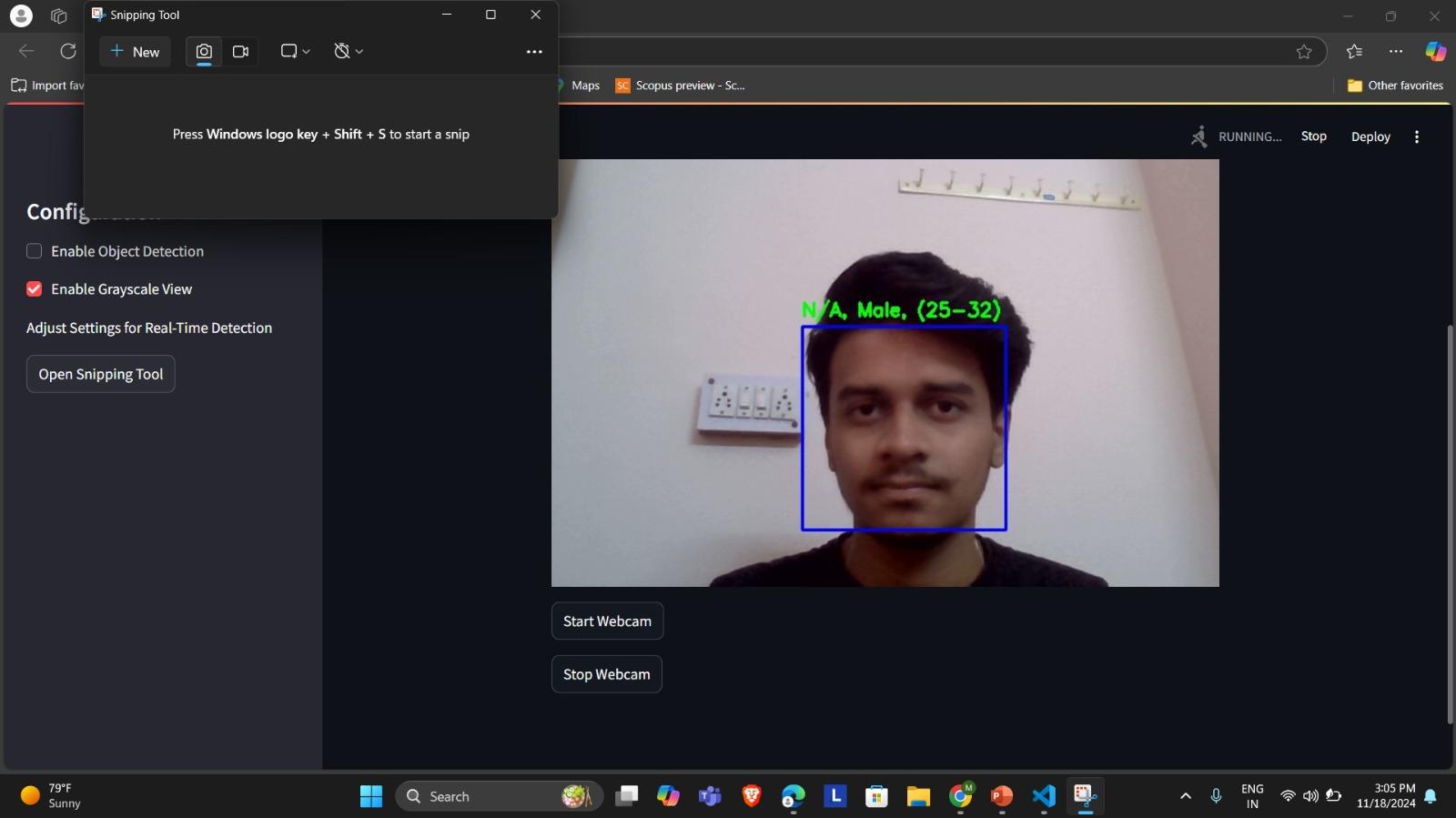


Fig. 5 ‘Open Snipping Tool’ option, which allows to take the screenshot of the results displayed.

IV. COMPARISON WITH OTHER METHODOLOGY

| **Sr No.** | **Comparison of Document Forgery Detection Techniques and their Performance** | | |
| --- | --- | --- | --- |
| **Paper** | **Methodology** | **Accuracy** |
| 1 | Facial Emotion Detection Using Deep Learning [10] | CNN-based architecture for emotion detection | 70.14% (FERC-2013)  98.65% (JAFFE) |
| 2 | Affective Computing Technology Review [11] | Automatic classifier for emotion detection using various sensors | N/A |
| 3 | Facial Emotion Recognition using CNN (FERC) [12] | Two-part CNN: background removal, facial feature extraction | 96% |
| 11 | Facial Expression Recognition using CNN [13] | Ensemble of CNNs, preprocessing with face landmark extraction | 75.2% |
| 24 | CNN-10 for Facial Emotion Recognition [15] | CNN-10 model optimized for emotion classification | 84.3% (FER-2013)  95.4% (JAFFE)  99.9% (CK+) |

IV. CONCLUSION AND FUTURE SCOPE

The system is able to demonstrate real-time emotion recognition, age and gender prediction, and object detection. It successfully proves the capability of computer vision combined with deep learning techniques to analyze live video streams with a great level of accuracy. Using the more complex models, such as CNN for the analysis of a face and YOLOv8 for object detection, it achieves high accuracy in time with respect to handling multiple different tasks. This has critical implications for a broad range of practical applications, including smart surveillance, interactive entertainment, targeted marketing, and personalized user experiences-in each case, the ability to comprehend human behavior and interactions in real-time is critical. In the future, many enhancements can be devised to improve the capabilities and performance of the system. Expansion of the added feature would improve the model's capabilities to include more attributes, such as ethnicity, facial landmarks, or facial action units, and perform a more sophisticated analysis of human emotions and expressions. The useful value would then accrue to more subtle insight in human-computer interaction and psychological research. System optimization further towards cross-platform deployment could extend the usability of the applications to edge devices, also enable real-time processing on mobile platforms, IoT systems, or even other highly resource-constrained environments. Ethical issues are involved in the design and development of AI-driven systems because these interact with human data. Hence, future developments on this line should focus on data privacy, reducing biases in the models, and complying with the ethics standards so that such misuse of such systems does not spread further. Working within such ethics, the system would promote responsible application and ensure trust from the users particularly in sensitive fields like health and law enforcement. [9]

In conclusion, the system comprises a powerful foundation toward real-time applications of emotion recognition and object detection. Building on the work presented here, the further iterations can be used to propagate the system into broader applications, higher accuracy, and great versatility with a very wide scope of industries

References

1. Khan, Amjad Rehman. "Facial emotion recognition using conventional machine learning and deep learning methods: current achievements, analysis and remaining challenges." Information 13.6 (2022): 268.
2. Dada, Emmanuel Gbenga, et al. "Facial Emotion Recognition and Classification Using the Convolutional Neural Network‐10 (CNN‐10)." Applied Computational Intelligence and Soft Computing 2023.1 (2023): 2457898.
3. Huang, Zi-Yu, et al. "A study on computer vision for facial emotion recognition." Scientific Reports 13.1 (2023): 8425.
4. Ali, Md Forhad, Mehenag Khatun, and Nakib Aman Turzo. "Facial emotion detection using neural network." the international journal of scientific and engineering research (2020).
5. Jaiswal, Akriti, A. Krishnama Raju, and Suman Deb. "Facial emotion detection using deep learning." 2020 international conference for emerging technology (INCET). IEEE, 2020.
6. Mellouk, Wafa, and Wahida Handouzi. "Facial emotion recognition using deep learning: review and insights." Procedia Computer Science 175 (2020): 689-694.
7. Mehendale, Ninad. "Facial emotion recognition using convolutional neural networks (FERC)." SN Applied Sciences 2.3 (2020): 446.
8. Singh, Dilbag. "Human emotion recognition system." International Journal of Image, Graphics and Signal Processing 4.8 (2012): 50.
9. Truong, Vinh. "Textual emotion detection–A systematic literature review." (2024).
10. *Diwan, Tausif, G. Anirudh, and Jitendra V. Tembhurne. "Object detection using YOLO: Challenges, architectural successors, datasets and applications." multimedia Tools and Applications 82.6 (2023): 9243-9275.*
11. Khorasani, Mohammad, Mohamed Abdou, and J. Hernández Fernández. "Web Application Development with Streamlit." Software Development (2022): 498-507.