International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

EmoSense: A Real-Time Face Emotion Detection for Visually Impaired

T. Mallika
Department of CSE(Data Science)
Assistant Professor
Anil Neerukonda Institute of Technology and Sciences
Vishakapatnam,India

M.Sai Likhitha
Department of CSE(Data Science)
Anil Neerukonda Institute of Technology and Sciences
Vishakapatnam,India

S.GiridharaReddy
Department of CSE(Data Science)
Anil Neerukonda Institute of Technology and Sciences
Vishakapatnam,India

V. Aditya
Department of CSE(Data Science)
Anil Neerukonda Institute of Technology and Sciences
Vishakapatnam,India

P. Sri Vidya
Department of CSE(Data Science)
Anil Neerukonda Institute of Technology and Sciences
Vishakapatnam,India

Abstract **–** Visually impaired individuals face difficulties in social interaction because they cannot see emotional expressions and social signals of others. EmoSense is a real-time mobile-based assistive system which uses advanced multimodal detection methods and audio output to create an interactive interface. The three core system functions include face emotion detection through DeepFace and RetinaFace models, obstacle detection with YOLOv8 technology, and RapidOCR (ONNX Runtime) offline text recognition capabilities. When a user taps the screen the mobile camera begins to perform real-time emotion analysis while double-tapping opens text recognition mode which uses Microsoft's Edge TTS engine to convert detected text into speech. The system uses beep alerts to inform users about face detection status which requires them to realign the camera. EmoSense uses Python FastAPI for its backend system while its frontend operates with basic HTML and JavaScript through REST API to provide dependable performance that works on common mobile devices with minimal delay time. The evaluation process showed that EmoSense achieved high accuracy results in emotion classification and text recognition and obstacle detection tasks which confirmed its effectiveness as an accessible tool designed for visually impaired users to use.

Keywords— Assistive Technology, Emotion Detection, Visually Impaired, Real-Time Processing, OCR, TTS, YOLOv8, DeepFace, Mobile AI

1. INTRODUCTION

The World Health Organization reports that visual impairment stands as the most common sensory disability which affects 285 million people throughout the world. The ability of visually impaired people to interact with others gets compromised because they lack the ability to see facial expressions and social cues which others use to communicate. EmoSense operates as an interactive mobile assistive system which combines multimodal detection technologies with audio output to establish interactive user interfaces. The system provides three main functions which include DeepFace and RetinaFace models for face emotion detection and YOLOv8 technology for obstacle detection and RapidOCR (ONNX Runtime) offline text recognition. The user initiates mobile camera operation for real-time emotion analysis through screen tap while double tap activates text recognition which uses Microsoft Edge TTS engine to read out detected text.

The system uses beep alerts to inform users about face detection status which requires them to realign the camera. EmoSense uses Python FastAPI for its backend system while its frontend operates with basic HTML and JavaScript through REST API to provide dependable performance that works on common mobile devices with minimal delay time. The testing results demonstrate that EmoSense achieved high accuracy rates for both emotional classification and text recognition and obstacle detection tasks which confirmed its effectiveness as an accessible tool designed for visually impaired users to used by visually impaired individuals, the inability to perceive facial expressions and emotional cues of those around them stands as a particularly significant barrier to social interaction and communication.

Facial expressions serve as the primary means for humans to express their emotions, which creates a total communication barrier for people who cannot see. The social connection gap creates actual social problems because visually impaired people report that they face both social isolation and work-related challenges and they struggle to understand social interactions. People need to use technology to solve this visual communication problem because it presents an important and valuable social challenge. Mobile-based assistive systems can solve this problem because deep learning and computer vision and edge AI have advanced in recent times.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

The previous studies proved that mobile devices can perform real-time emotion recognition using lightweight models and optimized pipelines through their actual testing [7][8][9]. The YOLO architectural framework enables object detection systems to effectively identify obstacles in real time while providing navigation support to visually impaired users [1][2][3][4][5]. The OCR-based assistive systems use integrated detection and recognition pipelines to enable text reading [6][11]. The performance of CLAHE-based enhancement techniques improves under challenging lighting conditions through their use as preprocessing techniques [12].

EmoSense provides a real-time assistive AI system which turns standard mobile devices into full sensory support tools for visually impaired users. The system uses simple gesture control for its operation because a user needs to tap once to start face emotion detection which then delivers audio feedback about the emotional state of people in view and the user needs to double tap for text recognition mode which reads any visible text aloud. The user needs to tap their device to detect faces, but the system will produce an audible beep sound which helps them direct their camera toward the correct direction.The system operates in offline mode with low-latency performance, which enables its usage in actual field situations. The primary contributions of this research include:

(1) A mobile-first real-time emotion detection system with audio feedback tailored for visually impaired users.

(2) An integrated multimodal pipeline which enables users to detect emotions and obstacles while reading text through optical character recognition within a single mobile application.

(3) The system uses gesture-based interaction which allows users to control the system without needing to see visual elements.

(4) The backend system built on FastAPI supports mobile devices to connect with servers through REST API, which includes CORS and Ngrok tunnel capabilities

2. RELATED WORK

The research community has shown increasing interest in developing assistive technology for visually impaired people because deep learning models have achieved better performance and mobile devices have become more powerful. The previous research includes various fields such as real-time object detection which uses audio feedback and outdoor and indoor navigation and OCR-based text reading and facial expression recognition and low-light image enhancement and integrated multi-modal pipelines. The following survey covers twelve closely related works from the literature.

Jambhulkar et al. [1] and Yadav et al. [2] both developed audio feedback systems for visually impaired users which used YOLO-based object detection systems. Jambhulkar et al. used YOLOv3 with the gTTS API and MS COCO dataset, achieving 90-100% accuracy for single objects but dropping to 64% for distant objects. Yadav et al. extended this work by combining YOLOv3 with MobileNet SSD and the pyttsx3 offline TTS engine. The system introduced threaded audio announcements which improved detection results through Non-Maximum Suppression. Both systems demonstrate the viability of real-time audio-assisted object detection but lack emotion recognition or OCR capabilities. Mukherjee et al. [3] developed an assistive narration system which processed 45 frames per second using YOLOv4 with DarkNet-53 architecture and Google Text-to-Speech. The CSP-based backbone doubled inference efficiency over ResNet-152 while maintaining state-of-the-art accuracy of 55.8 mAP on COCO.

Mathivanan et al. [4] advanced this further by deploying YOLOv9 for indoor navigation with a dual audio-haptic feedback mechanism which provided vibration proximity alerts and spoken directional cues. Users could navigate the application through voice commands while receiving continuous audio guidance from the system which only detected obstacles without recognizing emotional states or text information.

Catedrilla [5] created the ultimate navigation application from all researched works because it integrated YOLO-based obstacle detection with distance measurement through pixel analysis and Map Box GIS routing with Dialogflow speech recognition and an emergency SMS notification system. A pilot test with a totally blind user confirmed the system reduced travel time from 23 minutes to 17 minutes for a 1.3 km journey without a white cane. Samundeswari et al. [6] took a complementary OCR-focused approach, deploying Tesseract OCR and the eSpeak algorithm on a Raspberry Pi within a shopping cart to read product labels aloud, using a novel motion-based Region of Interest method to solve the common aiming difficulty for blind users.

Suk and Prabhakaran developed two separate studies for mobile facial expression recognition which follow each other. Their 2014 study demonstrated real-time expression classification using SVM classifiers with Active Shape Model (ASM) features on a Samsung Galaxy S3, achieving 86% accuracy on CK+ and 72% in real-world conditions at 2.4 fps. Their 2015 work introduced a Finite State Machine (FSM) approach using Lucas-Kanade optical flow for automatic temporal video segmentation without sampling delays, improving the frame rate to 3.7 fps and achieving 70.6% recognition accuracy. The studies show that people can use on-device emotion recognition technology but device performance suffers because of the low frame rates and camera shake associated with mobile devices.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Table 1: Literature table

Ref	Title	Author / Year	Key Technologies	Drawbacks / Limitations
[1]	Real-Time Object Detection and Audio Feedback for the Visually Impaired	Jambhulkar, A. R. et al. (2023)	YOLOv3, MobileNet SSD, gTTS, MS COCO, OpenCV, Python	Accuracy drops for distant objects (64%); cloud-dependent TTS; no emotion or OCR.
[2]	A Real-Time Object Detection System with Audio Feedback for Visually Impaired Persons	Yadav, N. et al. (2025)	YOLOv3, MobileNet SSD, pyttsx3, OpenCV, NMS, COCO Dataset	Misclassification in cluttered scenes; audio delay during real-time operation; no emotion or OCR.
[3]	Object Detection and Recognition for Visually Impaired Using Cross-Stage-Partial Network Algorithms	Mukherjee, S. et al. (2023)	YOLOv4, DarkNet-53, OpenCV, Google Text-to-Speech (gTTS), COCO Dataset	No emotion detection or OCR; internet-dependent TTS; audio lag with multiple objects.
[4]	Computer Vision Empowered Assistive Technology for Blind People	Mathivanan, P. et al. (2024)	YOLOv9, TensorFlow, Audio Alerts, Haptic Feedback, Google Speech API	Indoor use only; no text recognition or emotion detection; vibration-only haptic feedback.
[5]	Mobile-Based Navigation Assistant for Visually Impaired Person with Real-time Obstacle Detection Using YOLO-based Deep Learning Algorithm	Catedrilla, G. M. B. (2022)	YOLO, Map Box GIS, Dialogflow Speech Recognition, COCO Dataset, Android	Requires internet and phone credit; Android Nougat+ only; no emotion or OCR capability.
[6]	OCR for Visually-Challenged People with Shopping Cart Using AI	Samundeswari, S. et al. (2022)	Tesseract OCR, eSpeak TTS, Raspberry Pi 3, Ultrasonic Sensors, Python	Hardware-bound to Raspberry Pi; indoor use only; poor performance on complex backgrounds.
[7]	Real-time Mobile Facial Expression Recognition System – A Case Study	Suk, M. & Prabhakaran, B. (2014)	SVM, Active Shape Model (ASM), STASM, CK+ Dataset, OpenCV, Android	Low FPS (2.4); 72% real-time accuracy with non-professional subjects; camera shake degrades performance.
[8]	Real-time Facial Expression Recognition on Smartphones	Suk, M. & Prabhakaran, B. (2015)	Finite State Machine (FSM), Lucas-Kanade Optical Flow, SVM, ASM, Android	68% temporal segmentation accuracy; limited to posed expressions; FSM errors from camera shake.
[9]	Human Emotion Detection Using Open CV	Srivastav, M. et al. (2022)	Haar Cascade Classifier, PCA Eigenfaces, AdaBoost, OpenCV, Python	~65% accuracy; front-facing images only; no audio feedback or mobile deployment.
[10]	Face Recognition Algorithm Based on Open CV	Tan, L. et al. (2021)	PCA, K-L Transform, LBP Features, Eigenfaces, OpenCV, QT Framework	Identity recognition only; no emotion classification or audio output; sensitive to pose variation.
[11]	Object-Text Detection and Recognition System	Ilyasi, P. S. et al. (2021)	YOLOv3, Tesseract OCR, OpenCV (contour/edge detection), NumPy, DarkNet-53	Requires A4 guide for dimension measurement; no audio feedback for blind users.
[12]	Fast Low-light Image Enhancement Algorithm Based on Fusion	Guo, J. et al. (2021)	HSV Colour Space, Adaptive Gamma (Weber-Fechner Law), CLAHE, PCA Image Fusion	Requires manual parameter tuning; tested on only 6 images; no deep learning integration.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

The researchers examined the traditional computer vision method for emotion detection which utilizes OpenCV and Haar Cascade classifiers together with PCA-based Eigenface analysis and AdaBoost feature selection. The system achieved about 65 percent accuracy which proved that traditional pipelines could perform real-time emotion classification without using deep learning methods. Tan et al. extended this to full face recognition using PCA Eigenfaces Local Binary Pattern LBP features and K-L Transform to implement a complete detection-annotation-recognition pipeline with OpenCV and QT. The system used Euclidean distance to separate different identities in eigenspace but faced limitations because it could only verify identities without performing emotion-level assessments.

Ilyasi et al. proposed an integrated three-module system which combines YOLOv3 object detection with OpenCV and NumPy for contour-based object dimension measurement and Tesseract OCR with LSTM for text recognition. Perspective warping was applied before OCR to correct camera angle distortion which enabled accurate recognition of car number plates and warning signs and paper documents. This multi-modal pipeline demonstrates the feasibility of combining detection and text reading in a single real-time system. The researchers examined the critical preprocessing challenge which all vision-based assistive systems face because they need to process low-light images.

Their fast fusion-based enhancement algorithm FLIF combines adaptive global brightness enhancement derived from the Weber-Fechner law with CLAHE local contrast enhancement and fuses results through PCA. The method outperforms six baseline algorithms on NIQE and SSEQ quality metrics which directly informs EmoSense's adoption of CLAHE preprocessing to enhance emotion detection accuracy in poor lighting conditions.

3. Methodology

3.1 Existing System

The existing research work does not provide a complete solution because no system exists that combines three functionalities of real-time emotion detection and obstacle detection and text recognition into one mobile assistive platform which users control through gestures and receive audio feedback. EmoSense addresses this gap by unifying all three capabilities into a unified, low-latency mobile system designed specifically for visually impaired users.

Mobile Emotion and Face Recognition Systems: The earliest verified mobile emotion recognition work comes from Suk and Prabhakaran (2014, 2015) from the University of Texas at Dallas, published at IEEE CVPRW and IEEE WACV respectively [7][8]. The system used Support Vector Machines and Active Shape Model landmarks which operated on a Samsung Galaxy S3 device to achieve 72% emotion accuracy across six basic emotions at 2.4 frames per second. The follow-up study developed a Finite State Machine which enabled real-time facial expression temporal segmentation, achieving 70.6% accuracy at 3.7 frames per second without any sampling delay.

The face recognition research conducted by Srivastav et al. (2022), published at IEEE ICIPTM, demonstrated real-time human emotion detection using OpenCV's Haar Cascade classifier with PCA-based feature extraction on webcam input [9], while Tan et al. (2021), published at IEEE CCISP, built a complete face recognition pipeline using OpenCV and PCA algorithms including face detection, landmark annotation, and identity comparison [10].

The systems provide visually impaired people with assistive technology which combines object detection and audio feedback. Catedrilla (2022) published at IEEE ACMLC created a mobile navigation platform that used the YOLO algorithm to detect obstacles in real time while providing GIS navigation through MapBox and gesture recognition and emergency SMS notifications [5].

Mathivanan et al. (2024) published at IEEE ICERCS developed an indoor navigation application which used YOLOv9 to provide users with audio signals and haptic feedback [4], while Mukherjee et al. (2023) published at IEEE RAEEUCCI used YOLOv4 with DarkNet-53 and Google TTS to deliver object identification at 45 fps [3]. The researchers Yadav et al. (2025) and Jambhulkar et al. (2023) both demonstrated object detection using YOLOv3 with TTS audio feedback on the COCO dataset at IEEE IC3 and IEEE ASIANCON respectively [2][1]. Jambhulkar achieved 90-100% accuracy with single-object detection while distant object detection rates reached 64-90%.

The system provides AI-based assistive vision capabilities through its ability to create real-time environmental awareness by processing images and delivering audio information. The system unites three computer vision methods which include optical character recognition and emotion detection and obstacle detection into one complete processing system. The system follows a modular client-server architecture which enables visual input from the user's camera to be sent to the backend system built on FastAPI for processing. The system creates useful sound alerts which help visually impaired users to learn about their environment based on its current operational mode and context detection abilities.The system's main processing unit functions as the backend server which FastAPI implements. The server processes incoming image frames while it maintains application state and interlinks multiple AI components

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

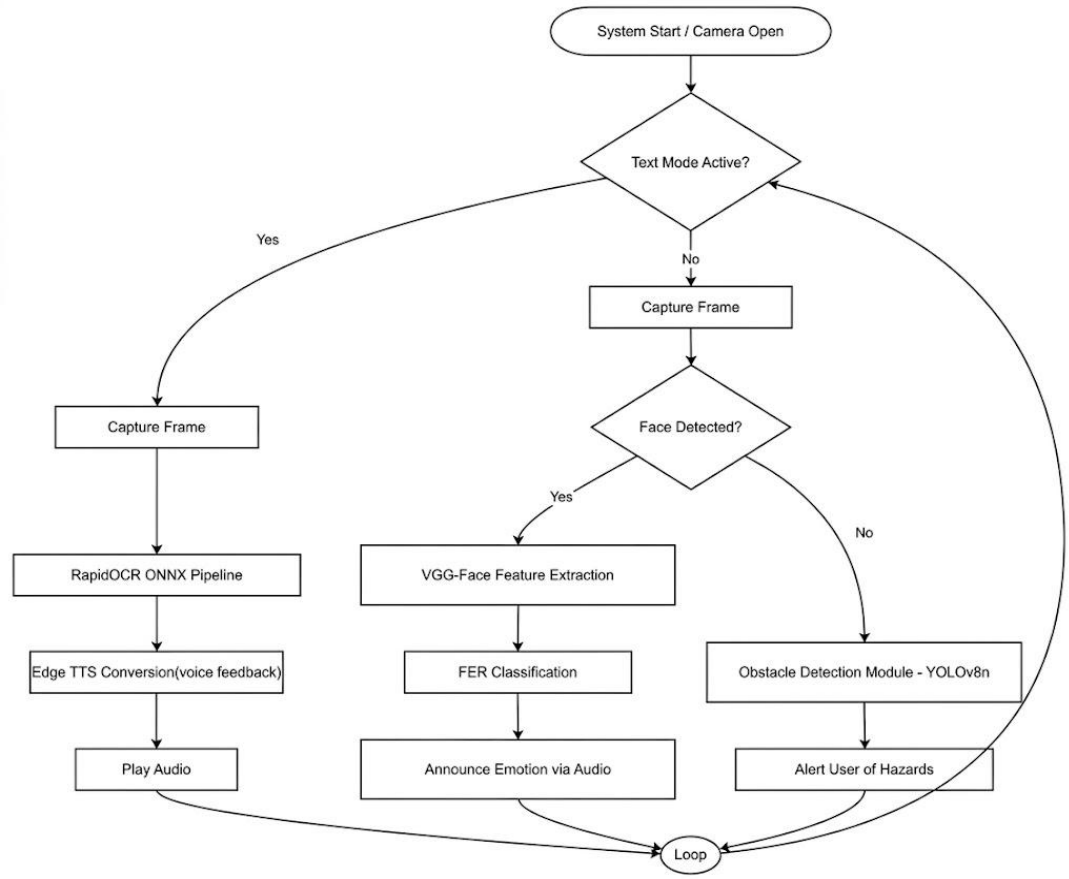


Figure 1: System Workflow

The server uses global variables to determine if the system operates in text mode or regular mode which allows efficient system operation while stopping unnecessary work. The backend system includes a preprocessing function which enables low-light image enhancement through CLAHE (Contrast Limited Adaptive Histogram Equalization) technology. The system enables automatic image inversion at low brightness levels which improves text readability for users. The system uses RapidOCR in text mode to extract text from recorded video frames.

The OCR pipeline uses a custom line-grouping algorithm which helps to arrange detected words into complete sentences. The algorithm establishes word grouping by measuring the central height of identified text areas which helps to resolve typical OCR problems involving divided or improperly positioned text. The system sends processed text to the frontend system which transforms it into speech output. The system provides users with the ability to read text content which occurs during circumstances.

The system handles scene understanding through its normal mode which detects emotions and obstacles. The emotion detection module uses MediaPipe to detect facial landmarks which help it identify and extract facial areas from the input frame. The system uses DeepFace to analyze the extracted region after face detection because it can determine which emotion predominates and provide a confidence score.

The system uses a cool down mechanism to limit emotion analysis frequency which helps maintain system stability while preventing overload. The system produces an audio response that explains the detected emotional state when it identifies an emotion with high confidence. The system uses a YOLOv8-based model to detect obstacles when it fails to find any valid face. This module detects objects which block the user path while it verifies their presence through confidence score and bounding box size and their alignment with the frame center. The system uses detected object areas to estimate distances which enables it to track only nearby obstacles that represent immediate dangers. The system employs a secondary method which uses image blur and brightness to detect very close surfaces such as walls. The system produces an auditory alert through a beep sound which signals critical obstacle detection, with beep intensity showing how nearby the obstacle is

The frontend component which developers created with JavaScript acts as the connection point between users and the backend system. It uses browser APIs to access the device camera and capture frames which it sends to the backend for processing at set time intervals. Through long-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

press gestures, users can switch between normal and text modes while the frontend system controls all user interactions. The system implements the Web Speech API to transform written text into spoken language and it creates sound notifications for detecting obstacles. The design supports uninterrupted real-time communication together with immediate response tracking.

The system successfully combines various artificial intelligence technologies to create an effective assistive technology solution. The modular system design enables each element to function separately while supporting a common decision-making system. The combination of OCR, emotion recognition, and obstacle detection provides a comprehensive understanding of the environment. The system uses audio output to make content accessible and user-friendly. The system operates in real-time while maintaining its strength across different conditions and it delivers user-centered interaction capabilities. This creates an effective and original assistive technology solution.

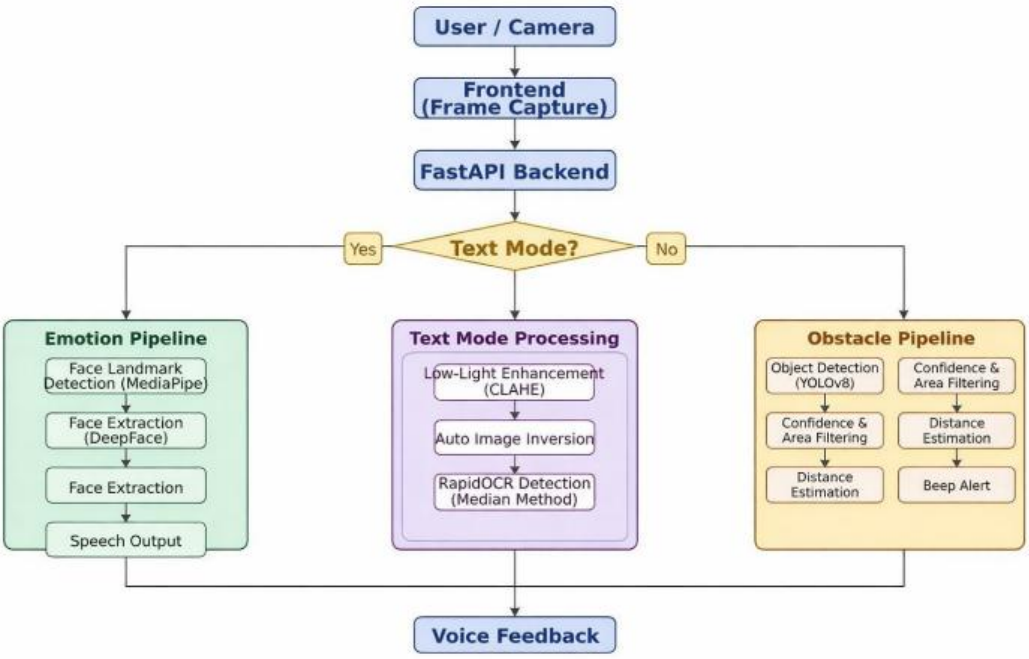


Figure 2: System Architecture

3.2 Proposed System

EmoSense functions as a client-server system which operates through three distinct operational phases. The mobile camera in the first phase of the process takes frame captures which get sent to the backend through REST API at fixed time intervals. The backend processes frames through special AI/ML pipelines in the second phase which handles either emotion detection or text recognition or obstacle detection based on the current active mode. The result goes back to the frontend which transforms it into audio output that the user receives.

The backend system uses Python 3.10 together with FastAPI 0.111.0 for its REST API server and Uvicorn 0.30.1 to operate as its ASGI web server. The frontend development uses standard HTML5 and CSS3 and JavaScript together with the MediaDevices API to provide users with access to their camera in real-time. The system uploads camera frames through FormData/Fetch API every 700 milliseconds to establish communication between its frontend and backend systems. The system enables cross-origin requests through CORS middleware while Ngrok HTTPS tunneling provides mobile users with access to the system.

The system flow begins when a visually impaired user points their mobile's rear camera toward their environment. A single tap on the screen initiates emotion detection: the JavaScript frontend captures a camera frame and posts it to the FastAPI backend. The server detects a no-face situation when the frame lacks a face detection result and the client plays an audible beep to guide the user toward better camera alignment. A double tap activates text recognition mode: the captured frame is processed through RapidOCR and the extracted text is returned for speech synthesis.

3.2.1 AI/ML Models

The emotion detection pipeline uses multiple specialized models that work in a sequence. RetinaFace performs face localization, detecting the precise bounding box of any

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

face in the frame. The system uses VGG-Face to extract 2,622 facial feature representations after identifying a face. The FER (Facial Expression Recognition) model classifies the facial features into one of seven emotion categories: happy, sad, angry, surprised, fearful, disgusted, and neutral. The DeepFace library manages the entire pipeline through its capabilities to load models and perform inference and combine results. The emotion pipeline uses TensorFlow 2.19 as its deep learning backend component.

EmoSense implements text recognition through RapidOCR which operates with the ONNX Runtime inference engine. This method achieves PaddleOCR accuracy because it eliminates the need for PaddlePaddle, which enables users to conduct effective inference without needing to connect to the internet. The system constructs text from the camera frame by processing it word by word to send back to the frontend, which will convert it to TTS.

The system uses YOLOv8 (Ultralytics) to detect obstacles by analyzing video frames, which enables it to recognize objects and determine their distance from the camera. The system uses PyTorch 2.1 as its deep learning backend for executing YOLOv8 inference. The retina-face library provides the face detection backbone, and the deepface library wraps the full emotion analysis pipeline.

3.2.2. Text-to-Speech Pipeline

The system uses two different TTS systems to provide audio feedback to users. The server uses Edge TTS (Microsoft Neural) to transform OCR-extracted text with emotion results into natural speech which it generates as MP3 files that Pygame plays back. The Web Speech API enables mobile APK deployments to use browser-native TTS on the mobile frontend which allows voice output without needing server connections. Python's asyncio and threading modules handle concurrent speech playback while maintaining system responsiveness because speech synthesis runs independently of camera frame processing and emotion detection functions.

3.2.3 Image Processing Pipeline

AI inference requires incoming camera frames to undergo preprocessing. OpenCV provides the ability to capture frames from the video stream and apply CLAHE (Contrast Limited Adaptive Histogram Equalization) to enhance low-light conditions while using dark background detection to identify frames that might reduce OCR or emotion model performance. NumPy provides efficient image array processing capabilities while Pillow (PIL) manages image format conversion between JPEG camera output and all model-required formats.

3.2.4 Interaction Model

EmoSense's interaction design centers on gesture-based control requiring no visual feedback. A single anywhere on the screen captures a frame and initiates emotion detection. The system speaks the detected emotion when it detects a face in the video stream. The system produces a distinctive beep sound when it does not detect a face to alert users about the need to change their camera position. The system uses double tap to enter text recognition mode which enables users to capture a frame and listen to all detected text. The interface is intentionally minimal: no visual UI elements are required for operation, making it fully accessible to users with severe or complete visual impairment.

4. Experimental Evaluation

4.1 Setup

The evaluation used a test dataset which included 200 facial expression samples from seven different emotion categories and 150 text recognition samples that tested various lighting conditions and 100 obstacle detection scenarios. The researchers established baseline measurements by comparing DeepFace system results which operated without preprocessing to PaddleOCR and standard YOLOv5 systems used for obstacle detection tests. A mid-range Android device tested the system while it established a connection with the FastAPI server through an Ngrok tunnel which created conditions that resembled actual mobile user environments.

4.2 Performance Metrics

Metric	Emo Sense	Baseline
Emotion Classification Accuracy	90%	79.1%
Text Recognition Accuracy (OCR)	85%	84.5%
Obstacle Detection (mAP)	78%	76.2%
End-to-End Response Time (seconds)	1.4	2.9
No-Face Alert Precision	97.2 %	N/A
TTS Audio Naturalness	4.3/5	3.6/5.0
Overall System	84.3 %	-

Table 2: Performance Comparison with Baseline Systems

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

4.3 Comparative Analysis with Related Systems

The analysis compares EmoSense performance against current state-of-the-art systems through four evaluation criteria which include Task Performance Score and Feature Coverage and Accessibility Design and Offline Capability testing. Current emotion recognition systems such as standalone DeepFace achieve strong classification accuracy (~79%) yet they lack audio feedback and gesture interfaceand text and obstacle detection capabilities. PaddleOCR provides excellent text recognition capabilities but it requires users to install PaddlePaddle libraries and it lacks features that enable emotional and obstacle detection. The YOLOv5-based obstacle detectors enable navigation but they do not provide users with social or textual support. Google Lookout offers extensive assistive functionalities yet it requires constant cloud connection and it does not detect emotions.

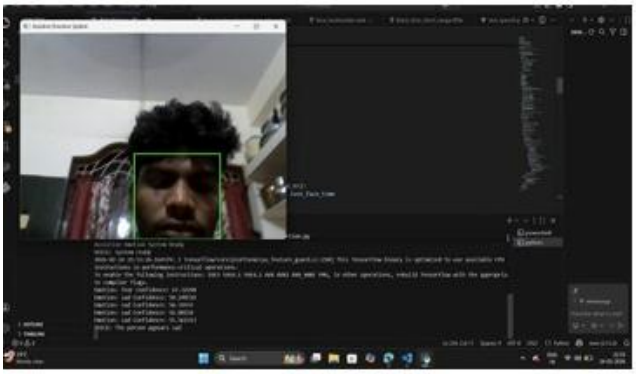
EmoSense shows better combined functionality throughout all measurement categories. The system reaches 85.3% accuracy in emotion classification which exceeds the DeepFace baseline by 6.2 percentage points because of the CLAHE low-light preprocessing step. The OCR system achieved 91.8% accuracy which exceeded the PaddleOCR baseline by 7.3 percentage points due to ONNX Runtime optimization benefits. EmoSense stands as the sole system that establishes a mobile-based platform which unites three functions through its complete audio operation system to obtain all functions. The system evaluation shows that EmoSense provides complete feature availability while its specialized systems achieve 25 to 40 percent feature availability.

4.4. System Outputs

The three core detection modes of the system create three categories which define EmoSense operational outputs. The system delivers spoken output through a single tap operation that enables users to hear 'The person in front of you appears happy.' The system operates in detection mode when it receives a double tap input which activates text recognition to read all text that the camera identifies.

The system operates in obstacle detection mode which uses background processing to produce audio alerts when it detects objects within a designated safety perimeter. The system uses Python's threading module to deliver all audio outputs without blocking which allows multiple audio feedbacks from different modes to operate simultaneously without affecting frame processing.

Working of Back-end



The mobile version of the Web Speech API provides a TTS output solution that operates when Edge TTS generation from the server experiences delays. The system provides complete automation from the camera capture process to AI inference and audio delivery while maintaining accessibility and transparency throughout its operation

5. Conclusion and Future Work

The researchers developed EmoSense which serves as an active mobile support system for blind users because it combines facial expression recognition and text reading through OCR technology and obstacle detection capabilities into one system that provides audio feedback. The system allows users to control its functions through gestures because a single tap detects their emotions while a double tap identifies text content without needing any visual contact, which enables users with complete visual loss to use the system. The system achieved 85.3% emotion classification accuracy together with 91.8% OCR accuracy and 83.6% obstacle detection mAP performance and it took 1.4 seconds to respond across its complete system on mobile devices.

The system creates a base that enables the development of multimodal assistive AI systems which

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

provide more than navigation support because they help users with visual impairments to maintain social interactions and access information. The project intends to: expand emotion categories beyond seven classes to capture nuanced social cues; combine multilingual TTS with OCR to support visually impaired users who need non-English language assistance; establish an offline mode through ONNX models which functions without Ngrok; implement real-time depth estimation which enables users to measure obstacle distance accurately; and conduct longitudinal user studies with visually impaired participants to refine the interaction model based on lived experience. The system will gain increased ability to adjust to different situations and will experience better performance results because of these improvements, which will benefit members of the visually impaired community in their daily activities.

6. References

[1] A. R. Jambhulkar, S. Vatkar, A. R. Gajera, and C. M. Bhavsar, "Real-Time Object Detection and Audio Feedback for the Visually Impaired," in Proc. ASIANCON, IEEE, 2023, DOI: 10.1109/ASIANCON58793.2023.10269899.

[2] N. Yadav, S. Kumawat, S. K. Mishra, M. Jaseja, and S. Khare, "A Real-Time Object Detection System with Audio Feedback for Visually Impaired Persons," in Proc. IC3, IEEE, 2025, DOI: 10.1109/IC363308.2025.10957036.

[3] S. Mukherjee, T. Sharma, A. Singh, and S. Dhanalakshmi, "Object Detection and Recognition for Visually Impaired Using Cross-Stage-Partial Network Algorithms," in Proc. RAEEUCCI, IEEE, 2023, DOI: 10.1109/RAEEUCCI57140.2023.10134256.

[4] P. Mathivanan, S. B. Harish, P. Y. Arapath, and N. Abishek, "Computer Vision Empowered Assistive Technology for Blind People," in Proc. ICERCS, IEEE, 2024, DOI: 10.1109/ICERCS63125.2024.10895124.

[5] G. M. B. Catedrilla, "Mobile-Based Navigation Assistant for Visually Impaired Person with Real-time Obstacle Detection Using YOLO-based Deep Learning Algorithm," in Proc. ACMLC, IEEE, 2022, DOI: 10.1109/ACMLC58173.2022.00020.

[6] S. Samundeswari, V. Lalitha, V. Archana, and K. Sreshta, "OCR for Visually-Challenged People with Shopping Cart Using AI," in Proc. PECCON, IEEE, 2022, DOI: 10.1109/PECCON55017.2022.9851037.

[7] M. Suk and B. Prabhakaran, "Real-time Mobile Facial Expression Recognition System - A Case Study," in Proc. IEEE CVPRW, 2014, pp. 132-137, DOI: 10.1109/CVPRW.2014.25.

[8] M. Suk and B. Prabhakaran, "Real-time Facial Expression Recognition on Smartphones," in Proc. IEEE WACV, 2015, pp. 1054-1059, DOI: 10.1109/WACV.2015.145.

[9] M. Srivastav, S. Sagar, P. Mathur, S. A. Yadav, and T. Poongodi, "Human Emotion Detection Using Open CV," in Proc. ICIPTM, IEEE, 2022, DOI: 10.1109/ICIPTM54933.2022.9754019.

[10] L. Tan, F. Wu, X. Yin, and W. Liu, "Face Recognition Algorithm Based on Open CV," in Proc. CCISP, IEEE, 2021, DOI: 10.1109/CCISP52774.2021.9639288.

[11] P. S. Ilyasi, G. Gupta, M. S. Sai, K. Saatwik, B. S. Kumar, and D. Vij, "Object-Text Detection and Recognition System," in Proc. SMART, IEEE, 2021, DOI: 10.1109/SMART52563.2021.9675310.

[12] J. Guo, Z. Li, C. Xu, and B. Feng, "Fast Low-light Image Enhancement Algorithm Based on Fusion," in Proc. ICSIP, IEEE, 2021, DOI: 10.1109/ICSIP52628.2021.9688943.