IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

BIONIC EYE MODEL TO PROVIDE VISION OR RESTORE SIGHT FOR BLINDNESS USING VISION TRANSFORMER

[1]Mrs.Vijaya Lakshmi D M, [2]Nishandhini A, [3]Prathika L, [4]Sangeetha S

[1]Assistant Professor, [2]Student, [3]Student, [4]Student

[1]Department of Computer Science and Engineering,

Adhiyamaan College of Engineering, Hosur, India.

Abstract: **:** This research addresses the challenges faced by visually impaired individuals in recognizing people, interpreting facial expressions, and engaging in social activities. Current visual implant systems, like the RPS, suffer from low resolution phosphene images, limiting their effectiveness in artificial vision solutions. To overcome these limitations, our project introduces a pioneering VT-based approach, adapting a deep learning architecture for natural image recognition. Our innovative method utilizes the VT to extract and process crucial information from the user's surroundings. By comprehending the visual environment, this system provides enhanced perception for visually impaired individuals, offering insights into count, familiarity, gender, estimated ages, facial emotions, surrounding objects, and approximate distances of individuals nearby. The integration of the VT into the artificial vision system aims to transcend the constraints of current technologies, providing visually impaired individuals with a transformative tool. This research not only contributes to the advancement of artificial vision but also has the potential to significantly enhance the quality of life for those with visual impairments.

Index Terms - Vision Transformer(VT),Retinal Prothesis System(RPS)

I. INTRODUCTION

Visual impairment poses a profound challenge to individuals, impacting their ability to navigate and engage with the world around them. Recognizing faces, interpreting facial expressions, and participating in social activities become formidable tasks for those with impaired vision. Despite the existence of visual implant systems such as the Retinal Prosthesis System, the limitations in the resolution of phosphene images have hindered the development of effective artificial vision solutions. In response to these challenges, our research endeavors to revolutionize artificial vision by introducing a Vision Transformer-based approach. This innovative strategy aims to overcome the constraints of existing visual implant systems and provide visually impaired individuals with an advanced and comprehensive visual perception tool. The Vision Transformer, originally designed for natural image recognition, offers a promising foundation for our endeavor. By adopting this deep learning architecture, we seek to extract valuable information from the user's surroundings, including crucial details such as the count, familiarity, gender, estimated ages, facial emotions, surrounding objects, and approximate distances of individuals. This holistic approach aims to address the diverse range of challenges faced by visually impaired individuals, offering them a transformative solution to enhance their engagement with the world. Through this research, we aspire not only to contribute to the evolution of artificial vision but also to make a meaningful impact on the lives of visually impaired individuals by providing them with a sophisticated tool that augments their perceptual capabilities. The integration of the Vision Transformer holds the potential to usher in a new era of artificial vision, where the limitations of current technologies are surpassed, enabling a more inclusive and enriching experience for those facing visual impairments.

II. LITERATURE SURVEY

Ghezzi **[1]** proposed a system called The Role of the Visual Field Size in Artificial Vision, visual field size in artificial vision and proposed that improving this aspect can enhance mobility and perform visually-driven tasks. The challenge of creating rudimental sight within a large VF. Epiretinal and suprachoroidal devices are discussed as options covering a significant portion of the retina, corresponding to a large VF size. Shruthi C. Poojary **[2]** proposed a system called A Review on Bionic Vision Technology, which provide blind or visually impaired individuals with the ability to see again. Including retinal implants, cortical visual prostheses, optogenetics, gene therapy, and stem cell transplantation. Implied that the technology utilizes retinal implants, cortical visual prostheses, optogenetics, gene therapy, and stem cell transplantation as methods to restore vision. XuanThanh-An **[3]** proposed a system called Artificial Vision: The Effectiveness of the OrCam in Patients with Advanced Inherited Retinal Dystrophies, the OrCam My Eye 2.0 in improving the quality of life and addressing rehabilitation needs in patients with advanced stages of RP or CRD. Patients with RP or CRD diagnoses and a best-corrected visual acuity of ≤20/400 Snellen were invited to participate. The OrCam My Eye 2.0 is not explicitly mentioned in the provided information. Shruthi C. Poojary **[4]** proposed a system called Blind Patients in End Stage Inherited Retinal Degeneration: Multimodal Imaging of Candidates for Artificial Retinal Prosthesis. cross-sectional study involves the review of clinical data and multimodal imaging of 40 eyes from 21 blind institutional patients with end-stage IRD. optical coherence tomography (SD-OCT), fluorescein angiography, and fundus autofluorescence for the analysis of the imaging features. Lorenzo Iuliano, and Giovanni Fogliato **[5]** proposed a system called Blind Patients in End Stage Inherited Retinal Degeneration: Multimodal Imaging of Candidates for Artificial Retinal Prosthesis, which Characterizes the imaging features of blind patients with end-stage IRD. blind institutional patients with end-stage IRD. The patients were screened for eligibility for the Alpha AMS retinal prosthesis. spectral-domain optical coherence tomography (SD-OCT), fluorescein angiography, and fundus autofluorescence for the analysis of the imaging features.

III. METHODOLOGY

Model Architecture

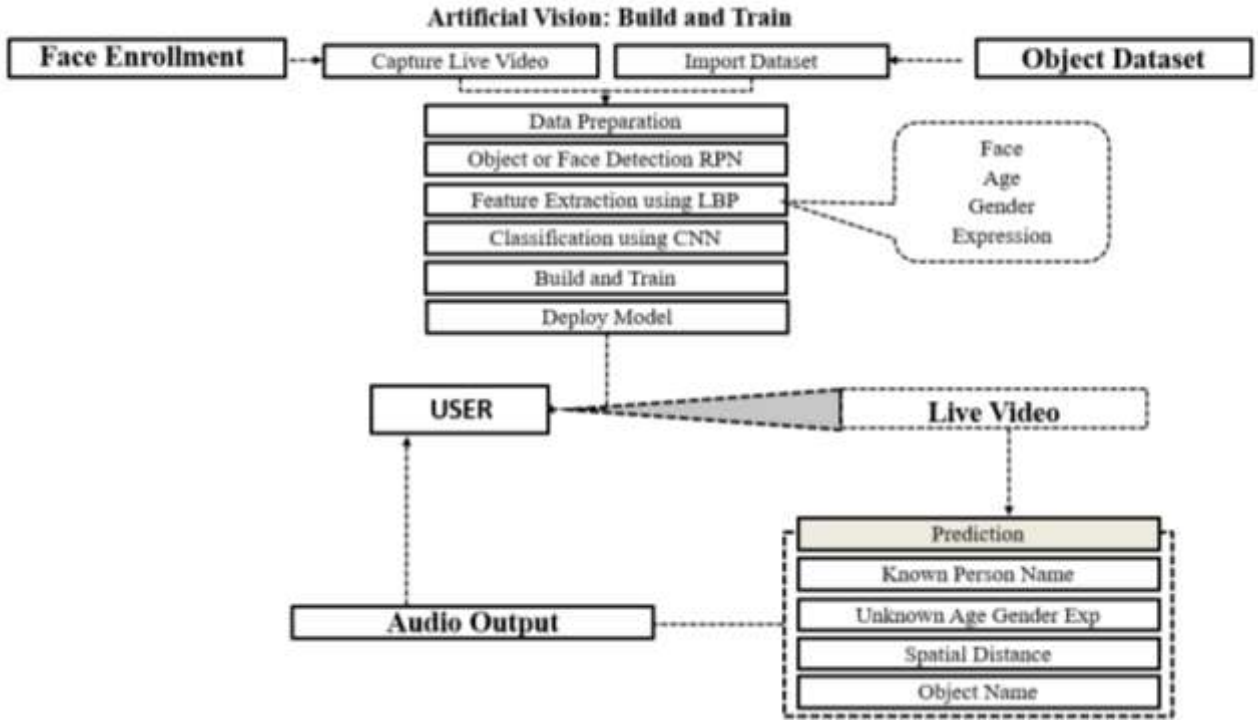

Figure 1: Architecture design of the model

Visual Implant Firmware: The Visual Implants Web App is designed to empower visually impaired individuals by leveraging cutting-edge technology. On the front end, user interfaces are crafted for simplicity, employing HTML templates and Flask-Bootstrap for accessibility. The back end, driven by Python and Flask, orchestrates image processing, object detection, facial recognition, and distance estimation modules. MySQL serves as the database, storing user credentials, training data, and preferences. The app's core functionality

includes real-time scene analysis, distinguishing known and unknown individuals, identifying gender, estimating ages, recognizing facial emotions, and determining approximate distances from the user. This integrated system aims to enhance the artificial vision experience, providing meaningful insights into the surroundings for improved mobility and quality of life.

i)End User Interface Admin Interface Login Module: Administrators gain secure access through this module, providing essential authentication for system control. Dataset Training Module: This module empowers administrators to upload, annotate, and manage datasets, forming the foundation for continuous machine learning model enhancement. The inclusion of feedback mechanisms allows dynamic adjustments based on real-world scenarios.

ii)Visually Challenged or Blind People Interface Live Video Module: This feature captures and streams real-time video through the user's glasses, with the Image Processing Module enhancing visual clarity on the live feed. Predicted Result Module: Utilizing object detection, facial recognition, and distance estimation, this module offers valuable insights into the surroundings, including the number of people, their identities, emotions, and approximate distances.

Audio Output Module: Designed for accessibility, this module converts predicted results into spoken information, employing text-to-speech capabilities. It provides auditory cues for users, including feedback on system status and voice-based navigation for user controls. The combination of live video predicted results and accessible audio output ensures a comprehensive and user-friendly experience for individuals with visual impairments.

iii)Artificial Vision Build and Train Dataset Collection: The initial step in building an artificial vision model involves the comprehensive collection of a diverse dataset. This dataset should encompass images relevant to face recognition, gender prediction, age estimation, and emotion prediction. Ensuring diversity in age, gender, and emotions within the dataset is crucial for robust model training. To create a robust dataset for Face Recognition, Gender Prediction, Age Estimation, and Emotion Prediction, a systematic approach involving live video capture and frame extraction is employed.

Live Video Capture: Utilize a camera or webcam to capture live video footage. This ensures that the dataset reflects real-world scenarios, providing diverse and dynamic facial expressions. Frame Extraction: Convert the live video into individual frames to create a sequence of images. This can be achieved using video processing libraries or frameworks. Extract frames at a regular interval to capture a representative sample of facial expressions. Face Recognition Dataset: From the extracted frames, manually or automatically annotate the faces to create a labeled dataset. Ensure diversity in lighting conditions, angles, and facial expressions to enhance the model's robustness.

Gender Prediction Dataset: For gender prediction, label each face in the dataset with the corresponding gender. This binary classification task involves associating each face with either 'male' or 'female.'

Age Estimation Dataset: Annotate the dataset with age labels to create an age estimation task. This involves associating each face with its corresponding age group or numerical age, depending on the level of granularity desired.

Emotion Prediction Dataset: Label each face in the dataset with the corresponding emotion, such as 'happy,' 'sad,' 'angry,' etc. This creates a multi-class classification task for emotion prediction. By systematically collecting a live video dataset and extracting frames, you can create a rich and diverse dataset suitable for training models for Face Recognition, Gender Prediction, Age Estimation, and Emotion Prediction. This dataset serves as the foundation for building accurate and inclusive artificial vision models. **Pre-processing:** The pre-processing phase is critical for preparing the dataset for model training. It involves several steps, including grayscale conversion to simplify processing, resizing to standardize image sizes, noise filtering to enhance image quality, and binarization if necessary for specific tasks. These pre-processing techniques ensure a consistent and optimized input for subsequent stages. Segmentation: Segmentation involves the identification and extraction of facial regions from images. In this context, Region Proposal Network (RPN) or similar face detection techniques are employed to locate and extract faces from the preprocessed images. This step is essential for isolating the regions of interest for subsequent analysis.

iv)Feature Extraction Face Recognition: Facial features are crucial for effective face recognition. Feature extraction techniques, such as Local Binary Patterns (LBP) or Histogram of Oriented Gradients (HOG), are applied to capture discriminative features from facial images. These extracted features serve as the foundation for accurate face recognition.

Gender Prediction: Building a gender prediction model involves training a machine learning model, potentially utilizing a deep learning framework like TensorFlow or PyTorch. The model is trained on the pre-processed and segmented dataset to accurately predict the gender of individuals based on their facial features.

Age Estimation: Similar to gender prediction, age estimation involves training a model to predict the age of individuals based on facial features. This typically involves regression techniques and the development of a model trained on the preprocessed and segmented dataset.

Emotion Prediction: Emotion prediction utilizes Local Binary Patterns (LBP) or similar techniques for feature extraction, focusing on capturing facial textures related to emotions. A dedicated model is trained on the pre-processed and segmented dataset for accurate emotion prediction. Object Detection: In addition to facial analysis, Object Detection is incorporated for a more comprehensive understanding of the scene. This involves identifying and localizing objects within the images, contributing to a holistic artificial vision system.

Classification: To unify face recognition, gender prediction, age estimation, and emotion prediction into a cohesive framework, a Convolutional Neural Network (CNN) architecture is employed. This CNN is trained on the integrated dataset, encompassing diverse tasks and ensuring effective classification across multiple dimensions.

Build and Train the Model: The final stage involves building and training the integrated model. The dataset is split into training and testing sets, and the model is iteratively trained. Parameters are adjusted to optimize performance, ensuring that the model effectively captures the intricate relationships within the data.

v)Vision Prediction System Live Video Analysis: The system begins by capturing a live video feed through a camera mounted on the user's device. This real-time video serves as the input for the Video Vision Transformer, a sophisticated model capable of efficiently analyzing visual information and extracting meaningful features.

Predict Know or Unknown for the Blind: The extracted features are then compared with the trained model to predict whether individuals in the user's vicinity are known or unknown. This comparison helps the system determine the familiarity of people, aiding the visually impaired user in recognizing acquaintances.

Predict Gender, Age, and Expression of Unknown Person: In cases where an unknown person is detected, the Video Vision Transformer goes beyond basic recognition. It delves into predicting additional details such as the gender, age, and facial expression of the unknown individual. This comprehensive analysis enhances the user's understanding of their surroundings.

Estimate Distance using MiDaS Algorithm: To provide spatial awareness, the system employs the MiDaS (Monocular Depth Estimation in Adverse Scenarios) algorithm for distance estimation. This algorithm calculates approximate distances between the visually impaired person and both known and unknown individuals in the scene. The information contributes to the user's understanding of the spatial layout.

Object Detection: Object detection is integrated into the Vision Prediction System to identify and locate objects in the user's surroundings. The Video Vision Transformer efficiently processes the live video feed to recognize and categorize various objects, enhancing the overall situational awareness of the visually impaired user.

Audio Output: To convey the system's predictions and information to the visually impaired user, an audio output component is incorporated. The predictions, including familiarity, demographics, expressions, and distances, are converted into auditory cues using text-to-speech technology. This ensures a seamless and informative user experience, allowing the user to receive real-time feedback through sound.

IV. EXISTING SYSTEM

A Kinect-Based Navigation System is a portable system that uses a standard Kinect sensor, battery, and laptop/processor. It uses Simultaneous locating and mapping technology (SLAM) from Google's Project Tango for indoor positioning, allowing for centimeter-level accuracy. Virtual Haptic Radar, a combination of a three-dimensional model and an ultrasonic-based motion capture system, is another example. Moovit is a free tool for public transport management, while BlindSquare conveys the relative location of previously recorded POIs through speech. Lazzus is a paid application that coordinates GPS and built-in motion capture and orientation sensors to provide users with intuitive cues about POIs in the surrounding area. The Histogram of Oriented Gradients (HOG) is a feature descriptor used to detect objects and faces in image processing and computer vision techniques. The Single Shot Detector (SSD) method detects objects in images using a single deep neural network, combining predictions from multiple feature maps with different resolutions to handle objects of various sizes.

V. PROPOSED SYSTEM

The Kinect-Based Navigation System is a portable device that utilizes a standard Kinect sensor, battery, and laptop/processor. The battery powers the Kinect and CPU for 3 hours. Smart Cane Simultaneous locating and mapping technology (SLAM) from Google's Project Tango allows centimeter-level indoor positioning. Intel RealSense provides vision positioning solutions, with applications in commercial drones like Yuneec's

Typhoon H.The study utilized a depth camera and a server for SLAM processing, enabling six-degrees-of-freedom indoor location and obstacle detection capabilities.

Virtual Haptic Radar originating from Haptic Radar, is a representative example. It substituted its predecessor's IR sensors with the combination of a three-dimensional model of the surroundings plus an ultrasonic-based motion capture system worn by the user. The user was prompted to approach an area near an object, and warning vibrations were triggered accordingly.

Moovit It's a free, effective, and easy-to-use tool that offers guidance on the public transport network, managing schedules, notifications, and even warnings in real-time. The National Organization of Spanish Blind People (ONCE) recommends it as an asset for mobility tasks.

BlindSquare Its specifically designed for the BVI, this application conveys the relative location of previously recorded POIs through speech. The system utilizes databases from Foursquare and OpenStreetMap.

Lazzus a paid application, again designed for BVI users, which coordinates GPS and built-in motion capture and orientation sensors to provide users with intuitive cues about the location of diverse POIs in the surrounding area, even including zebra crossings. It offers two modes of operation: the 360° mode verbally informs of the distance and orientation of nearby POIs, whereas the beam mode describes any POI in a virtual field of view in front of the smartphone. The primary data sources for this information are Google Places and OpenStreetMap.

Vision-Based System Histogram of Oriented Gradients (HOG) Histogram of oriented gradients (HOG) is a feature descriptor that is utilized to detect objects and faces in image processing and other computer vision techniques. The Histogram of oriented gradients descriptor technique includes occurrences of gradient orientation in localized portions of an image, such as the detection window, and the region of interest (ROI), among others. HOG-like features offer a simplicity and ease of understanding of their information.

Single Shot Detector (SSD) Single Shot Detector (SSD) is a method for detecting objects in images using a single deep neural network. The SSD approach divides bounding box output space into a set of default boxes based on different aspect ratios. After discretizing, the method scales per feature map location. The Single Shot Detector network utilizes multiple feature map predictions with varying resolutions to effectively handle objects of varying sizes.

VI. RESULT

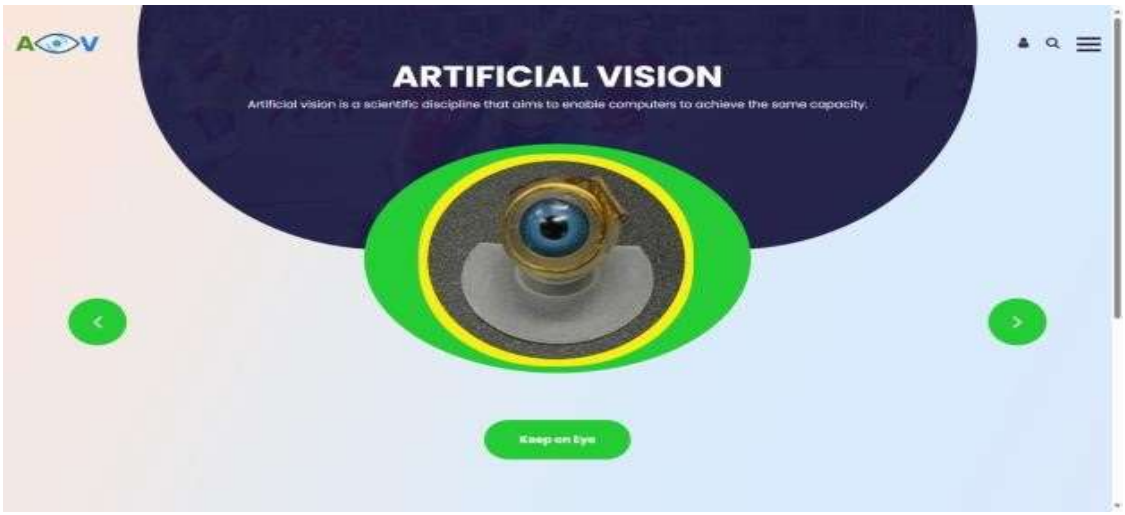

Figure 2: End User Interface

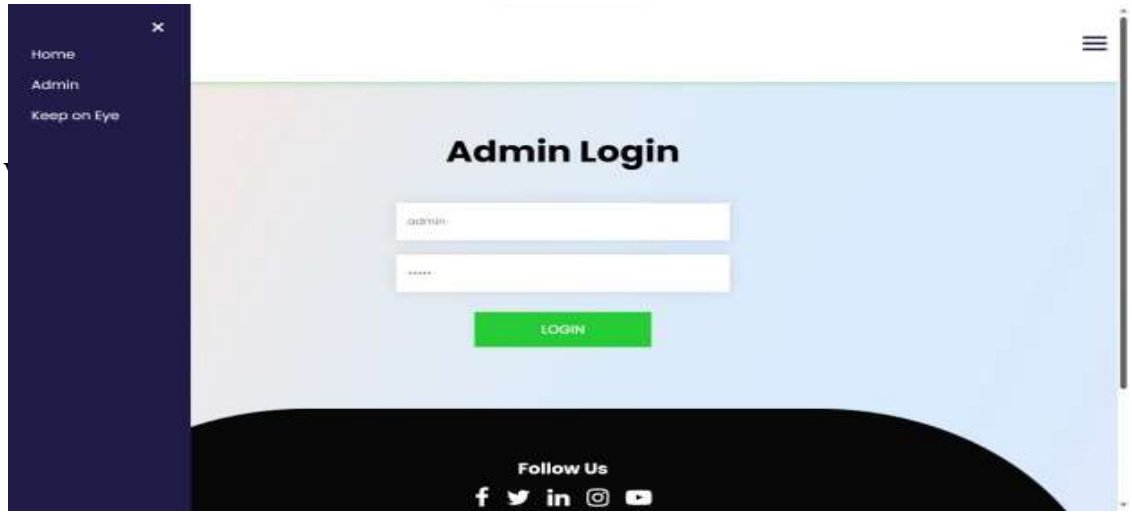

Figure 3: Admin Training Module

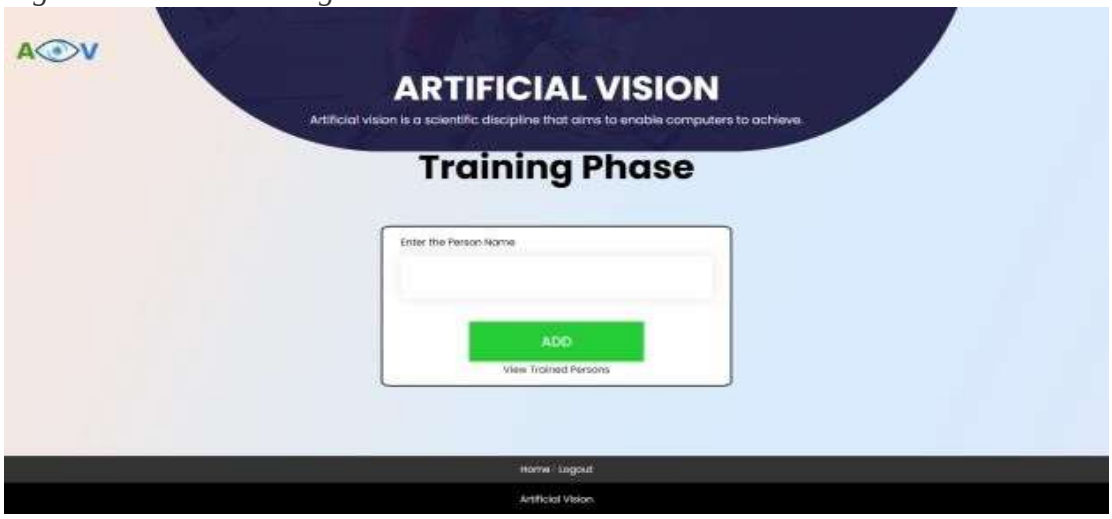

Figure 4: Model Training Interface

Figure 5: Admin Feedback and Iteration

Figure 6: Blind User Live Video Module

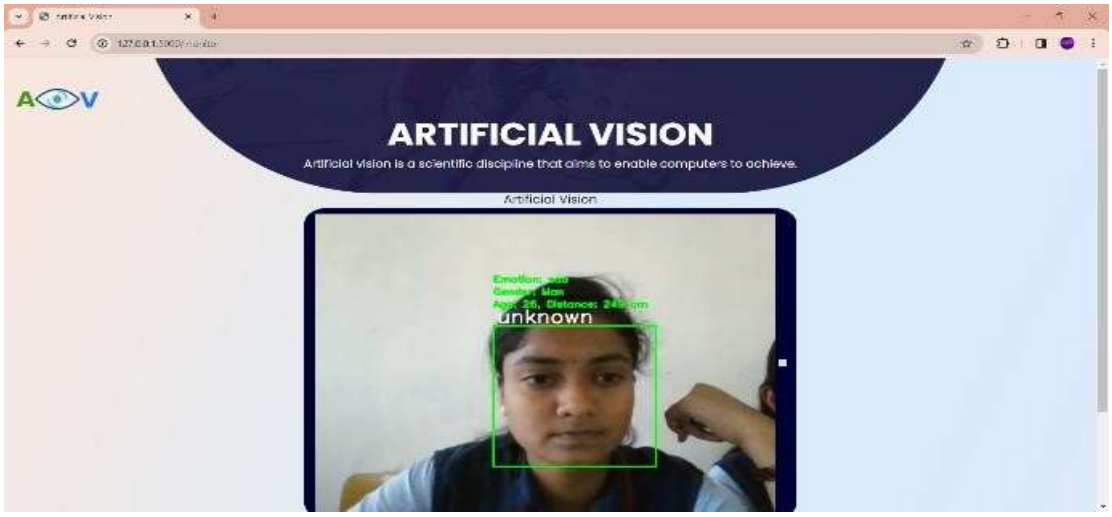

Figure 7: Vision Prediction System

VIII. CONCLUSION

In conclusion, the project aims to revolutionize the artificial vision experience for visually impaired individuals by integrating advanced technologies and innovative approaches. Through the implementation of Vision Transformer technology, real-time image processing algorithms, and information extraction techniques, the project endeavors to enhance accessibility, promote independence, and improve the overall quality of life for visually impaired users. By addressing key challenges such as limited access to information, navigation barriers, social interaction limitations, and educational and employment obstacles, the project seeks to empower visually impaired individuals to lead more fulfilling and independent lives. The integration of audio output with text-tospeech conversion ensures accessible feedback, while the validation with a simulated prosthetic vision and the feasibility analysis for everyday use further solidify the project's potential impact. This project not only aims to improve the daily lives of visually impaired individuals by providing a heightened artificial vision experience but also contributes to the broader field of artificial vision technologies. By fostering accessibility, independence, and user satisfaction, the project strives to set a benchmark for future innovations in visual implant systems. Moreover, by laying the groundwork for next-generation visual implant systems and exploring integration with existing visual implant technologies, the project sets the stage for continuous advancements in the field of artificial vision. Through user satisfaction assessments and feedback collection, the project aims to continually refine and improve the developed solutions to better meet the needs of visually impaired individuals.

IX. REFERENCES

[1] F. Catherine, Shiri Azenkot, Maya Cakmak, "Designing a Robot Guide for Blind People in Indoor Environments," ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts, 2015.

[2] H. E. Chen, Y. Y. Lin, C. H. Chen, I. F. Wang, "Blindnavi: a mobile navigation app specially designed for the visually impaired," ACM Conference Extended Abstracts on Human Factors in Computing Systems, 2015.

[3] K. W. Chen, C. H. Wang, X. Wei, Q. Liang, C. S. Chen, M. H. Yang, and Y. P. Hung, "Vision-based positioning for Internet- of-Vehicles," IEEE Transactions on Intelligent Transportation Systems, vol. 18, no.2, pp. 364–376, 2016.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[5] J. Ducasse, M. Macé, M. Serrano, and C. Jouffrais, "Tangible Reels: Construction and Exploration of Tangible Maps by Visually Impaired Users," ACM CHI Conference on Human Factors in Computing Systems, 2016.

[6] J. Engel, T. Schops, and D. Cremers, "LSDSLAM: Large-scale direct monocular SLAM," European Conference on Computer Vision, 2014.

[7] S. Gilson, S. Gohil, F. Khan, V. Nagaonkar, "A Wireless Navigation System for the Visually Impaired," Capstone Spring, 2015.

[8] J. Guerreiro, D. Ahmetovic, K. M. Kitani, and C. Asakawa, "Virtual Navigation for Blind People: Building Sequential Representations of the RealWorld," International ACM SIGACCESS Conference on Computers and Accessibility, 2017.

[9] Kendall, M. Grimes, and R. Cipolla, "Pose Net: a convolutional network for real-time 6-DOF camera relegalization," International Conference on Computer Vision, 2015.

[10] Kendall, and R. Cipolla, "Geometric loss function for camera pose regression with deep learning," International Conference on Computer Vision, 2017 measurement for population science", 2018.