International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1079-1085
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0152

Real-Time Vision Explainer: Object Recognition and Multilingual Voice Narration

Dr. A. Radhika[1], Mandapati Teja[2], Polukonda Durga Kalyani[3], Yalamanchili Grahita[4], Challa Deepa Malika[5]
[1]Professor, Department of Computer Science and Engineering, SRK Institute of Technology, Vijayawada,India
[2,3,4]Students, Department of Computer Science and Engineering, SRK Institute of Technology, Vijayawada, India
Emails: *radhikaankala@gmail.com[1], tejamandapati658@gmail.com[2], polukondadurgakalyani@gmail.com[3], yalamanchiligrahitha@gmail.com[4], Deepuchalla216@gmail.com[5]*

Abstract
Understanding visual information is a challenge for many users, especially when images contain multiple objects, people, or complex scenes. Existing solutions often provide limited explanations, depend heavily on proprietary systems, or lack multilingual and voice-based accessibility. This creates a gap for users who require detailed, understandable, and inclusive interpretation of visual content. To address this problem, this project proposes a real-time vision explanation system that analyzes images and provides meaningful explanations. The system allows users to authenticate securely and upload or capture images for processing. It detects objects and recognizes people using computer vision techniques, generates a descriptive summary of the image, translates the description into multiple languages, and produces voice narration for better accessibility. The system also stores user history in a structured text format for future reference. By combining vision analysis, language translation, and speech synthesis using free and open-source tools, the proposed solution enhances accessibility, learning, and interaction with visual data for a wide range of users.

Keywords: *Real Time Vision, Object Detection, Multilingual Translation, Voice Narration, Artificial Intelligence, Computer Vision.*

1. Introduction
The rapid growth of digital media and visual content has created a strong demand for systems that can effectively interpret and explain images in a meaningful way. Images often contain multiple objects, people, and contextual information that may not be easily understood by all users, particularly those with visual impairments or language barriers. While existing applications provide partial solutions, many of them rely on closed platforms, offer limited explanations, or lack multilingual and voice-based support. This highlights the need for an open, accessible, and intelligent vision-based system that can analyze visual content and present it in an understandable and user-friendly form. The proposed project, Real-Time Vision Explainer: Object Recognition and Multilingual Voice Narration, is designed to address these challenges by integrating computer vision, natural language processing, and speech synthesis into a single unified system. The project is structured into multiple phases, each responsible for a specific function, ensuring modularity, clarity, and ease of implementation. The first phase focuses on user authentication and system access, where users can register and log in securely using Firebase Authentication. This phase ensures that only authorized users can access the system and that individual user activities can be tracked for personalized history management. Authentication also improves data security and enables future scalability of the system. The second phase involves

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1079-1085
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0152

image acquisition, where users can upload an image or capture one directly using a camera interface. This phase serves as the input stage of the system and is designed to handle different image formats while validating the input for further processing. Proper image handling at this stage ensures smooth execution of subsequent analysis tasks. The third phase is dedicated to object detection and face recognition. Using computer vision techniques, the system analyzes the uploaded image to identify objects and recognize people present in the scene. This phase extracts essential visual features and forms the foundation for generating meaningful explanations. Accurate detection at this stage directly impacts the quality of the final output. The fourth phase focuses on description generation, where the detected objects and recognized faces are converted into a human-readable textual explanation. This phase bridges the gap between visual data and natural language, allowing users to understand the image content clearly without requiring technical knowledge. The fifth phase introduces multilingual translation, enabling the generated description to be translated into multiple languages. This phase enhances the usability of the system for users from diverse linguistic backgrounds and promotes inclusivity. By supporting multiple languages, the system becomes suitable for educational and assistive applications. The sixth phase implements voice narration, where the translated text is converted into audio output using text-to-speech technology. This phase significantly improves accessibility, especially for visually impaired users, by allowing them to hear the image description instead of reading it. The final phase manages result presentation and data storage, where the system displays the analyzed results to the user and stores the interaction history in a structured text or JSON format. This phase enables users to revisit previous analyzes and supports future enhancements such as analytics and personalization. By organizing the project into clearly defined phases, the proposed system ensures ease of development, maintainability, and extensibility. The integration of vision analysis, language translation, and speech output using free and open-source tools makes this project a practical, accessible, and scalable solution for real-time image understanding.

2. Literature Survey

1. Glenn Jocher works on YOLOv8 presents significant enhancements to the YOLO family of real-time object detectors. The paper focuses on improving detection accuracy, model robustness, and inference speed through an updated architecture and optimized training strategies. YOLOv8 introduces lightweight model variants suitable for edge devices while maintaining strong performance on standard benchmarks. Its contributions strengthen real-time object detection in practical applications, especially in scenarios requiring fast and efficient processing[1]. 2. Taigman et al propose DeepFace, a deep learning framework that reduces the performance gap between machine-based face verification and human-level accuracy. The system utilizes a nine-layer neural network combined with a 3D face alignment process that normalizes facial variations before recognition. Their approach significantly lowers face verification error rates and demonstrates the effectiveness of deep neural architectures in large-scale face recognition tasks[2]. 3. Bahdanau et al introduce an attention-based neural machine translation model that jointly learns to align source and target language representations. The attention mechanism allows the model to selectively focus on relevant parts of the input sentence during translation, resolving limitations of earlier encoder–decoder architectures. This work improves translation quality for long and complex sentences and establishes a foundation for modern multilingual translation systems[3]. 4. Xu et al introduce one of the first image captioning models that incorporates a visual attention mechanism, allowing the system to selectively focus on different regions of an image while generating each word of a caption. Their approach combines a convolutional neural network for image feature extraction with a recurrent neural network that dynamically attends to the most relevant visual features. This attention-based strategy results in captions that are more descriptive, accurate, and contextually aligned with image content, significantly improving over earlier fixed-feature

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1079-1085
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0152

models. The work laid the foundation for modern vision-language models and attention-driven captioning systems[4]. 5. Itseez Developers the OpenCV library, introduced by the Itseez developers, is a comprehensive open-source platform for real-time computer vision and image processing. The paper describes the library's wide range of functionalities, including image filtering, feature extraction, face detection, and object tracking. OpenCV is optimized for high performance, supports multiple programming languages, and runs across major operating systems, making it a widely adopted tool in academia and industry[5]. 6. Google AI the documentation for Google's Text-to-Speech Toolkit (gTTS) presents a multilingual speech synthesis system that converts textual input into natural-sounding audio. The toolkit supports various languages and dialects, offering clear and intelligible speech output for accessibility, virtual assistants, and interactive applications. Its lightweight Python interface makes it easy to integrate into real-time and multilingual systems[6]. 7. Davis King King's work on the face_recognition API, built upon the dlib machine learning framework, provides an accessible and accurate solution for face detection and recognition. The system utilizes Histogram of Oriented Gradients (HOG), convolutional neural networks, and deep metric learning to generate 128-dimensional facial embeddings for reliable identity matching. The API is designed for simplicity and practical deployment, making it popular for both research and production environments[7]. 8. Tan et al address the challenges of deploying object detection models on resource-constrained edge devices. The paper explores efficient neural network architectures, model compression methods, and optimization strategies that reduce computational load while preserving accuracy. Their work enables real-time inference on mobile and embedded platforms, supporting modern applications requiring low-power, on-device processing[8]. 9. Li et al present a vision–language pretraining framework that learns joint representations from large-scale image–text datasets. The model enables downstream tasks such as image captioning, visual question answering, and multimodal reasoning. Their work demonstrates how unified vision-language embeddings significantly enhance cross-modal understanding and helps form the basis for modern multimodal AI systems[9]. 10. Tiedemann work on surveys open-source and free translation APIs designed to support multilingual communication without relying on commercial platforms. The paper highlights lightweight neural translation models and public datasets that enable developers to integrate translation capabilities at no cost. This approach enhances accessibility and encourages the adoption of multilingual tools across diverse applications[10].

3. Ease of Use

3.1.User Authentication and System Access

Before utilizing the features of the proposed system, users are required to authenticate themselves through a secure login mechanism. The system employs Firebase authentication to manage user registration and login, ensuring that only authorized users can access the application. This authentication process provides a controlled environment where user activities can be securely managed and personalized. By implementing a structured access mechanism, the system enhances data security and supports reliable tracking of user interactions throughout the application lifecycle.

3.2.Image Input and Interaction Simplicity

The system is designed to provide a simple and intuitive method for users to interact with visual data. Users can upload an image from their device or capture an image directly using a camera interface. The image input module validates the selected image to ensure compatibility with the processing pipeline. This approach minimizes user effort and reduces errors during input selection, allowing a seamless transition to subsequent analysis stages. The simplicity of the input mechanism improves overall usability and ensures that users with minimal technical knowledge can operate the system effectively.

4. Related Work

4.1.Object Detection and Face Recognition Consistency

Once an image is submitted, the system performs

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1079-1085
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0152

object detection and face recognition using trained computer vision models. These models are configured with predefined parameters to maintain consistent performance across different inputs. By adhering to standardized detection thresholds and processing workflows, the system ensures reliable extraction of visual features. Maintaining consistency at this stage is essential, as the accuracy of detected objects and recognized individuals directly impacts the quality of the generated explanations.

4.2.Description Generation and Language Translation Accuracy

The detected visual information is transformed into a meaningful textual description that accurately represents the image content. This description follows a structured format to preserve clarity and relevance. The system then translates the generated text into multiple languages without altering the original meaning. By maintaining integrity during description generation and translation, the system ensures that users receive accurate and understandable information regardless of their language preference.

4.3.Voice Narration and Result Display

To enhance accessibility, the translated description is converted into voice narration using text-to-speech technology. This feature enables users, particularly those with visual impairments, to understand image content through audio output. The system presents both textual and audio results through a clear and organized interface, ensuring that information is easily accessible and user-friendly.

4.4.History Storage and System Reliability

All user interactions and analysis results are stored in a structured text or JSON format. This storage mechanism allows users to revisit previous analyzes while maintaining system reliability and data consistency. By preserving analysis history in a standardized format, the system supports future enhancements and ensures dependable long-term operation.

5. Existing System

The existing image search systems mainly rely on traditional techniques such as keyword-based search, metadata matching, and basic visual feature extraction. These systems depend heavily on manually added tags, file names, and textual descriptions, which often leads to inaccurate or incomplete results when images are not properly labeled. Older computer vision methods like color histograms, shapes, and texture analysis are also limited because they cannot truly understand the objects or meaning within an image. As a result, existing systems fail to recognize complex scenes, multiple objects, or semantic relationships, and they perform poorly when images vary in lighting, angle, or background. Additionally, these traditional approaches struggle to scale efficiently with large datasets, making real-time and highly accurate image search difficult.

6. Proposed System Architecture

Before implementing the proposed project, all required content and system functionality were planned and written clearly. The overall objective, system flow, and features were finalized before coding started. This approach helped in reducing errors and improving clarity during development. All text-related content was prepared separately, and images, datasets, and code files were maintained in organized folders. This ensured a clean structure and made the system easy to understand and manage.

6.1.Use of Terms and Abbreviations

All abbreviations and technical terms used in the project are defined clearly when they first appear in the document. Common terms related to computer vision and artificial intelligence are used consistently throughout the paper. Abbreviations are avoided in headings wherever possible to keep the content simple and readable. This helps readers easily understand the system without confusion.

6.2.Data Handling and Processing

The system processes images, text, and audio in a simple and structured manner. Images uploaded by the user are checked before processing to avoid errors. Object detection results are stored as text information, and translations are handled using standard language formats. Audio outputs are generated in common formats so that they can be played on any device. Using a consistent data format improves system reliability.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1079-1085
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0152

6.3.Processing Flow and Algorithm Usage

The system follows a sequential processing flow to then analyzed to detect objects and recognize faces. Based on the detected information, a human-readable description is generated. This description is translated into the selected language and converted into voice narration. Each processing stage operates independently, making the system easier to debug, maintain, and extend.

6.4.Common Design Practices

To avoid common mistakes, the system uses clear naming conventions and modular design. Each function performs a specific task, which improves readability and maintenance. Errors such as invalid image input or processing failure are handled properly with user-friendly messages. All user data and results are stored in a structured text or JSON maintain simplicity and clarity. After successful user authentication, an image is uploaded or captured and format, ensuring easy retrieval and future extension.

7. System Implementation

The system is implemented using Python and Flask for backend development. Firebase Firestone is used for structured data storage and authentication. AI-based question generation is powered by large language models integrated through APIs. Blockchain functionality is implemented using a custom cryptographic ledger to store hashes of approved questions. The frontend is developed using HTML, Tailwind CSS, and JavaScript, providing role-based access for administrators, faculty, and students. PDF generation is handled using Report Lab, ensuring standardized and secure document output.

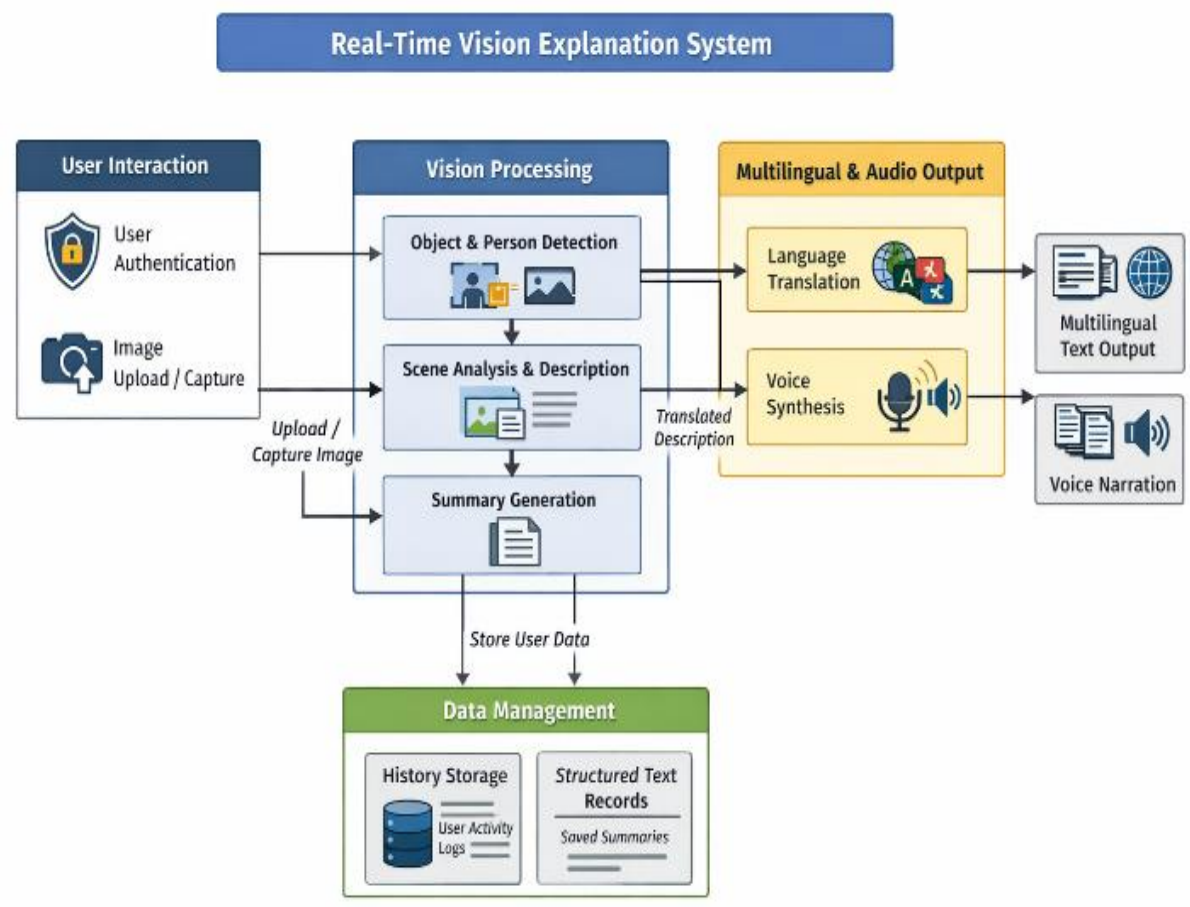

Figure 1 System Architecture

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1079-1085
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0152

8. Results and Discussion

The results of our project is given below.

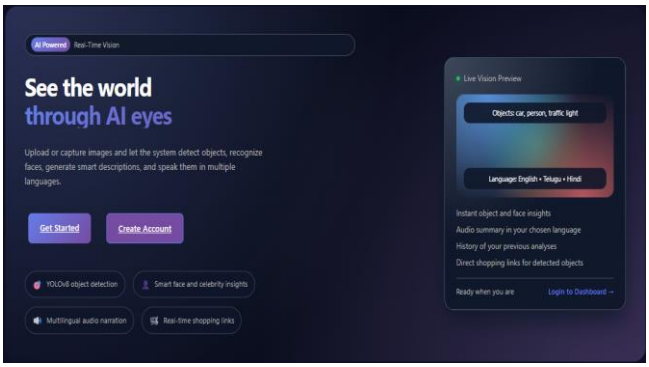

Figure 3 Welcome Page

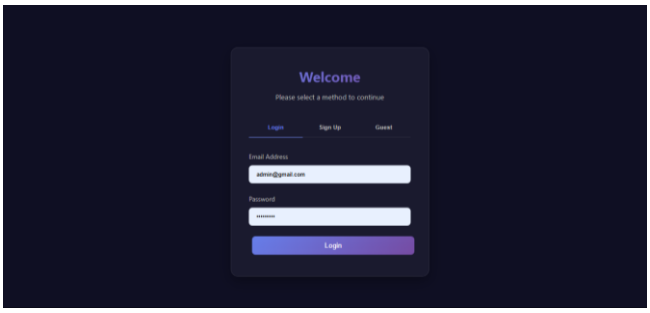

Figure 4 Login Page

Figure 5 User Dashboard

Figure 6 Trained Images

Figure 7 Trained Places

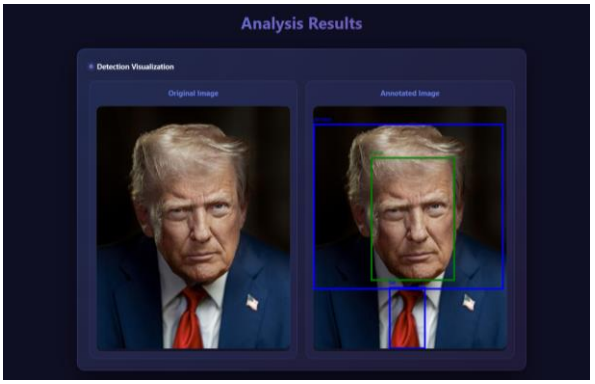

Figure 8 Image Search Analysis

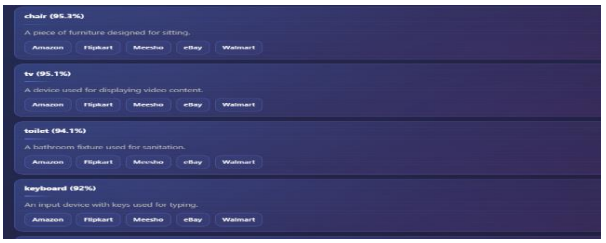

Figure 9 Detected Objects

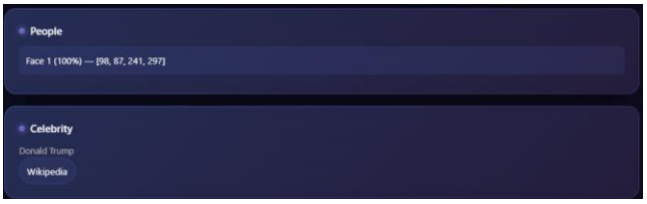

Figure 10 Person Detection

Figure 11 Multilingual Translation

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 03 March 2026
Page No: 1079-1085
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0152

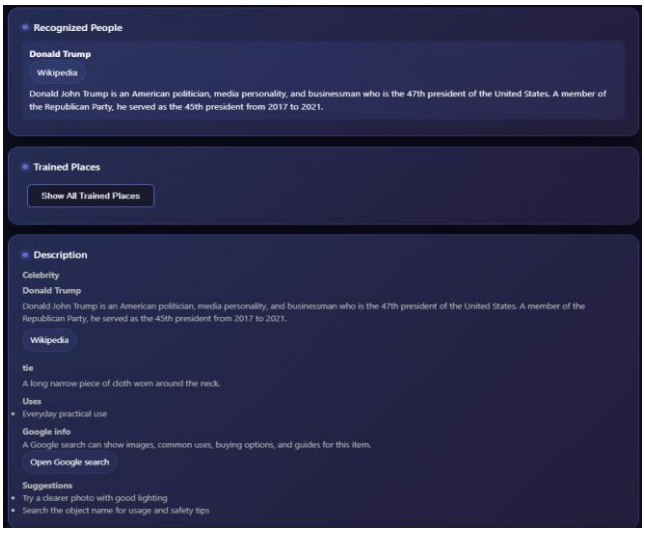

Figure 12 Detailed Analysis

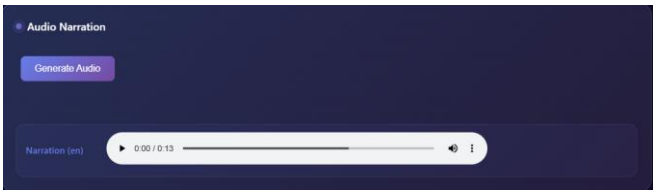

Figure 13 Audio Narration

Conclusion

The Real-Time Vision Explainer: Object Recognition and Multilingual Voice Narration system successfully integrates computer vision, natural language processing, and speech synthesis to provide an accessible and intelligent solution for understanding visual content. By combining object detection, face recognition, description generation, translation, and voice narration into a unified framework, the system delivers clear and meaningful explanations of images to users with diverse needs. The modular design ensures smooth authentication, efficient image handling, accurate visual analysis, and multilingual audio output, making the system highly user-friendly and inclusive. Through the use of open-source tools and lightweight architectures, the proposed solution remains scalable, cost-effective, and adaptable for real-time applications. The system's ability to store results and maintain structured user history further enhances usability and supports future personalization and analytical features. Overall, this project demonstrates a practical and impactful approach to enhancing digital accessibility, assisting visually impaired users, supporting educational environments, and improving interaction with visual information across multiple languages. Future extensions may include video support, advanced scene understanding, and real-time mobile deployment to expand the system's capabilities even further.

References

[1]. YOLOv8: Real-Time Object Detection Improvements" – Glenn Jocher
[2]. "DeepFace: Closing the Gap to Human-Level Face Verification" – Taigman et al.
[3]. "Multilingual Neural Machine Translation by Jointly Learning to Align and Translate" – Bahdanau et al.
[4]. "Automatic Image Captioning with Attention Mechanism" – Xu et al.
[5]. "OpenCV: Library for Real-Time Computer Vision" – Itseez Developers
[6]. "gTTS: Google Text-to-Speech Toolkit" – Google AI
[7]. "face_recognition API Based on dlib" – Davis King
[8]. "Real-Time Object Detection on Edge Devices" – Tan et al.
[9]. "Vision-Language Pretraining for Multimodal Understanding" – Li et al.
[10]. "Free Translation APIs for Cross-Language Communication" – Tiedemann