JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Speech Recognition for Visually Impaired Users

Nanda Hunagund [1], Dr. Meghana Kulkarni[2]

[1]M. Tech Student, Department of Electronics and Communication Engineering, Visvesvaraya Technological University, Belagavi -18, Karnataka, India.
Email: nandiniravitorgal18@gmail.com
[2]Professor Department of Electronics and Communication Engineering, Visvesvaraya Technological University, Belagavi -18, Karnataka, India.
meghanak@vtu.ac.in

Abstract

Individuals with visual impairments face continuous challenges when interpreting surrounding visual information, particularly in unfamiliar situations. To address this need, this work presents a system that automatically converts images into spoken descriptions in multiple languages. The system accepts visual input through a webcam or uploaded image file, generates captions with a pretrained BLIP model, translates the caption into the selected language, and converts the translated text into speech. The entire process requires minimal human interaction and supports Kannada, Hindi, Tamil, Telugu, French, and English. Experimental testing demonstrates that the system produces meaningful descriptions, correct translations, and clear speech output. By transforming visual content into sound, the system offers practical support for accessibility and independence among visually impaired users.

Keywords: speech processing, BLIP model, speech recognition

1. Introduction

Millions of individuals rely on assistive tools such as screen readers and voice assistants to access digital text. However, interpreting real-world images, surroundings, or printed content remains difficult without external help. Existing assistive applications often depend on single languages, cloud-only processing, or require manual intervention from caretakers. Building a system that generates natural descriptions of images and reads them aloud in different languages can reduce dependence on others and provide real-time situational awareness.

The aim of this project is to create a pipeline that can convert visual information into spoken language. The user either captures a scene using the system's webcam interface or uploads an image stored on the device. The BLIP model interprets the visual features and produces a descriptive sentence. The caption is translated to the chosen language and spoken aloud using text-to-speech synthesis. This allows visually impaired users to "hear" the content of images rather than rely on sight.

2. LITERATURE REVIEW

Research in assistive vision has evolved from basic object recognition to natural-language interpretation of complete scenes. Earlier works relied on handcrafted features and could recognize a small set of objects with limited accuracy. With the introduction of deep learning, models such as CNNs and Transformers have allowed richer feature extraction and contextual understanding. Vision-language frameworks like BLIP and CLIP combine image and text embeddings, enabling models to describe scenes in full sentences rather than labeling objects individually.

Studies on accessibility also highlight the importance of multilingual support. Many visually impaired users understand information better in their native language, making language flexibility essential in real-world applications. Recent systems combine captioning with speech synthesis, but many require constant internet connectivity or provide robotic-sounding speech. This project merges modern captioning performance with natural-sounding multilingual speech and real-time processing, creating a broader and more practical assistive solution.

3. SYSTEM ARCHITECTURE

The system follows a sequential architecture consisting of four main modules:

1. **Image Acquisition Module** – captures live frames from the webcam or loads stored images.
2. **Caption Generation Module** – uses the pretrained BLIP model to describe the image in English.
3. **Translation Module** – converts the caption to the desired language.
4. **Text-to-Speech Module** – produces speech output and plays it for the user.

Each stage passes its output directly to the next, forming a complete pipeline from image to audio. The architecture supports interactive use and runs on local hardware, making it suitable even when internet access is limited.

4. METHODOLOGY

The BLIP framework serves as the heart of the captioning process. The image is divided into visual patches and processed by the Vision Transformer. A short text prompt is tokenized by the text encoder, and both sets of features are aligned inside a multimodal encoder. The conditional decoder then produces a descriptive caption one word at a time.

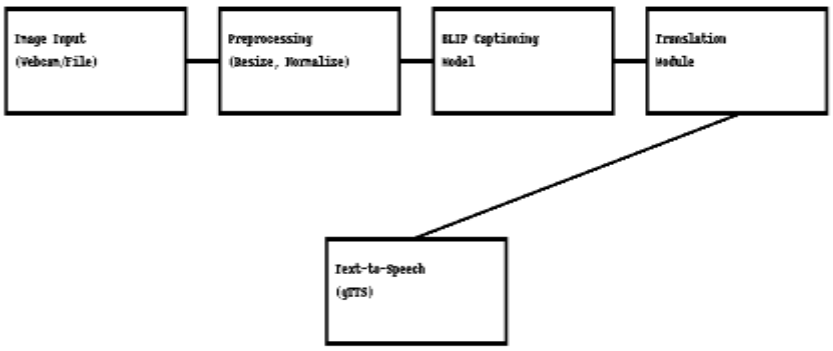

Figure 1: methodology of the work

For multilingual support, the generated caption is sent to the translation unit. The GoogleTranslator API ensures grammatical sentence formation and vocabulary adapted to each language. The translated text is sent to the speech synthesis module, where gTTS produces an MP3 audio file. After playback, temporary audio files are removed to maintain storage efficiency.

User interaction is handled through GUI and command-line interfaces. Buttons and keyboard controls allow users to capture images, upload files, select languages, listen to output, or exit the application. Error messages are provided when invalid paths or unsupported files are encountered.

5. IMPLEMENTATION

Python is used as the main programming language. OpenCV captures webcam frames and displays a live feed. The BLIP model and processor are loaded through the Hugging Face Transformers library. The caption generator runs entirely on the local machine, which avoids data-privacy concerns.

The translation module supports Kannada, Hindi, Tamil, Telugu, French, and English. The gTTS engine converts the translated caption into natural speech. Playback is handled automatically, and once audio output finishes, the temporary MP3 file is removed to prevent accumulated storage.

The GUI version built with Tkinter provides a simple interface where users can select languages and choose between webcam or file input. Icons, buttons, and pop-up windows provide feedback without requiring visual interpretation of text files.

$MathematicalModel$

$Let\ the\ input\ image\ be\ represented\ as$:

$[I \in R^{H\times W\times C}]$

$where\ (H), (W), and\ (C)\ denote\ height, width, and\ number\ of\ channels\ respectively.$

1. $ImageFeatureExtraction$

$The\ Vision\ Transformer\ in\ BLIP\ converts\ the\ input\ image\ into\ a\ sequence\ of\ feature\ vectors$: $[V = f_{ViT}(I)]$

$where$

$[\ V = \{v_1, v_2, \dots, v_n\}, \quad v_i \in R^d.\ ]$2. $TextPromptEncoding$

$Let\ the\ initial\ prompt\ be$:

$[T_0 =$ ``You are seeing a…"$]$

$The\ text\ encoder\ converts\ this\ prompt\ into\ embeddings$:

$[E = f_{text}(T_0)]$

$where\ [E = \{e_1, e_2, \dots, e_m\}, \quad e_j \in R^d.\ ]$

3. $MultimodalFusion$

$The\ multimodal\ encoder\ aligns\ visual\ and\ textual\ embeddings\ into\ a\ shared\ representation$:

$[M = f_{mm}(V, E)]$

4. $CaptionGeneration$

$The\ conditional\ decoder\ generates\ a\ caption\ token - by - token$:

$$[\ \hat{T} = \arg\max_{T} \prod_{k=1}^{K} P(\ t_k \mid t_1, t_2, \dots, t_{k-1}, M\ )]$$

$where(\hat{T})isthegeneratedEnglishcaptionand(K)isthecaptionlength.$

5. $TranslationModule$

$Given\ a\ target\ language\ (L), the\ translated\ caption\ is$:

$[\ \hat{T_L} = f_{trans}(\hat{T}, L)]$

6. $SpeechSynthesis$

$The\ translated\ text\ is\ converted\ to\ audio\ waveform\ by$:

$[A = f_{tts}(\hat{T_L})]$

7. $FinalSystemOutput$

$The\ final\ output\ consists\ of\ translated\ text\ and\ spoken\ audio$:

$$[Output = \{\hat{T_L}, A\}]$$

6. RESULTS AND DISCUSSION

Testing was performed with a range of real-world images containing people, objects, indoor scenes, and outdoor environments. The caption generator consistently produced relevant descriptions, even for images that were not part of the model's training data. Translations were accurate and retained the meaning of the English captions. Speech output was clear and intelligible.

Images captured in low light or with blurry backgrounds showed reduced caption detail, but the model still generated broad descriptions rather than producing errors. The system responded quickly, allowing near real-time

interaction. Users without technical knowledge were able to operate the interface successfully. The results confirm that the system can assist visually impaired users in understanding their environment independently.

Translation Module – Results and Discussion

The translation stage plays an important role in extending the system to users who are more comfortable with their native language. Once a caption is generated in English, the translation module converts it into one of the supported regional or international languages. During evaluation, the system produced correct and meaningful translations, and the wording reflected natural sentence structure rather than literal word-by-word conversion. For languages such as Kannada, Hindi, Tamil, Telugu, and French, the phrasing remained grammatically correct and easy to understand. This confirms that the translation component not only delivers linguistic accuracy but also preserves the intended meaning of the original caption, making the output understandable for a wider audience.

Text-to-Speech Module

After translation, the text-to-speech module converts the caption into spoken output. The gTTS engine produced clear speech with proper pronunciation, and the audio clips were generated in just a few seconds, even on standard hardware. To avoid unnecessary storage usage, each generated audio file was removed after playback. Continuous testing showed that the speech engine remained stable during repeated runs, without interruptions or audio glitches. Overall, the module proved lightweight, fast, and suitable for real-time use, allowing visually impaired users to receive instant voice feedback without needing to read text.

User Interaction and Control Module

User interaction is managed through a simple interface, allowing users to choose a preferred language, select either webcam or file input, and trigger caption generation. The system responded instantly to these selections, and the interface displayed helpful prompts and messages to guide the user. For example, if the image path was invalid or the file could not be processed, a warning message was shown. If the image was successfully handled, the caption and audio output appeared without delay. This level of responsiveness ensures that users—especially beginners or visually impaired individuals—can operate the system with minimal confusion and without needing technical knowledge.

Overall Discussion

Together, the translation, speech, and interaction modules form a complete pipeline that converts images into meaningful spoken output. The system functioned reliably across various test scenarios, including indoor and outdoor photographs, people, household objects, and daily activities. The combination of accurate captioning, smooth translation, and natural speech makes the tool useful for real-world accessibility applications. Although further improvements such as offline translation and customizable voices can be added later, the current implementation provides a stable and practical solution.

Input Captured Image – Analysis

The test image shows a woman seated indoors in front of a decorated wooden background. The lighting is bright and even, allowing details of her clothing, surroundings, and facial features to be captured clearly. The presence of household items and decorative elements in the background offers enough visual information for the captioning model to describe the scene accurately. This example demonstrates that the system can handle real-world images taken in natural lighting and produce meaningful descriptions.

Figure 2: Input image

Uploaded Image – Process Explanation

After an image is uploaded, the software stores it temporarily and displays it so that the user can confirm the input. When the "Generate Caption" button is pressed, the system forwards the file to the captioning module, which analyzes the visual content and returns a descriptive sentence. The interface then shows the caption to the user and plays an audio version of the same explanation. If any problem occurs—such as a corrupted image or unsupported file—the user receives a clear error message. When processing is complete, the temporary image is deleted to keep the system clean and efficient.

Caption Generation through GUI

If the user prefers the graphical interface, the system works in a similar way. When a file is selected from the file browser, the program attempts to read it and run the caption generator. The output is shown in a popup window, along with voice playback. This design removes the need for command-line typing and makes the application easy to operate for non-technical or visually impaired users. Error prompts ensure that improper inputs are handled safely, and valid images produce instant spoken descriptions.

Figure 3: Uploaded image

Live Image Through Webcam

When the webcam option is used, the program displays a live video feed and captures a frame when instructed. In the illustrated example, the system identified that the woman was looking at her mobile phone. When Kannada was selected as the output language, the model translated the caption and played the spoken result in Kannada. This feature is especially valuable for blind or low-vision users because it gives them an audible description of what is happening around them. By listening to the narration, users can understand objects, actions, and surroundings without needing to see the display.

Figure 4: live captured image with voice assistance

7. CONCLUSION

This Paper shows that visual scenes can be transformed into spoken language through an automated pipeline. By combining BLIP captioning, multilingual translation, and text-to-speech output, the system provides valuable assistance to visually impaired users. It reduces dependence on others, supports several languages, and works with both stored images and real-time camera input. With further enhancements, this system can evolve into a complete accessibility tool for education, mobility, communication, and daily life.

References

1. Khan M., Paul P., Rashid M., Hossain M., Ahad M. – "AI-based Visual Aid with Integrated Reading Assistant for Blind Users," 2020.

2. Bai J., Lian S., Liu Z., Wang K., Liu D. – "Wearable Navigation System for Blind Pedestrians Using Visual SLAM," 2019.

3. Chen X., Wang Y., Zhou F. – "Augmented Reality Based Navigation for Blind Users," 2018.

4. Joseph S., Banerjee R., Lee J. – "Social Media-Assisted Navigation for Visually Impaired," 2015.

5. Han J., Li P., Sun W. – "Human-Centric Video Understanding via Brain-Inspired Feature Mapping," 2021.