ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep311

PictoVoice — Pictures That Talk: An AI Based Image Captioning and Speech Generation System

Deepthi R. S.[1]; Bindushree U.[2]; Amrutha Raghupathy[3]; Shivani N.[4]

[1;2;3;4]Department of Information Science and Engineering SJB Institute of Technology Bangalore, India

Publication Date: 2026/09/18

Abstract: PictoVoice is an image-to-speech system designed to assist users in understanding visual information through spoken descriptions. The system accepts an image as input and processes it to extract relevant visual features, which are then used to generate a meaningful textual caption. The generated caption is subsequently converted into speech, enabling users to understand the contents of an image through audio output. The system integrates image processing, visual feature extraction, image captioning, and speech generation into a unified pipeline. Multilingual support for languages is currently being implemented to improve accessibility for users from diverse linguistic backgrounds.

Keywords: *Image Captioning, Image Processing, Visual Feature Extraction, Speech Generation, Text-to-Speech, Accessibility.*

How to Cite: Deepthi R. S.; Bindushree U.; Amrutha Raghupathy; Shivani N. (2026) PictoVoice — Pictures That Talk: An AI Based Image Captioning and Speech Generation System. *International Journal of Innovative Science and Research Technology*, 11(9), 308-313. https://doi.org/10.38124/ijisrt/26sep311

I. INTRODUCTION

Images are an important way of sharing information, but the visual content in an image may not be equally easy for everyone to understand. Identifying objects, people, and activities in a scene usually requires visual perception and interpretation. This creates a need for systems that can describe visual information in a form that is easier to access and understand.

Recent developments in deep learning have made it possible to analyze images and automatically generate descriptions of their content. Image captioning combines visual feature extraction with language generation to produce a meaningful sentence from an image. If this description is then converted into speech, the information can also be accessed through audio instead of relying only on visual or textual output.

PictoVoice – Pictures That Talk is developed with this idea in mind. The system takes an image provided by the user and processes it to identify the important visual information. The extracted features are used by the caption generation module to produce a textual description of the image. This description is then passed to a Text-to-Speech module so that the user can listen to the generated information.

The project brings image processing, visual feature extraction, caption generation, and speech generation together in a single workflow. ResNet-50 is used for extracting visual features, while an LSTM-based model is used to generate the caption from the extracted visual representation. The generated caption is finally converted into spoken output.

The current implementation focuses on completing the image-to-speech pipeline, from uploading an image to producing its corresponding audio description. Multilingual support is currently under implementation and is considered an ongoing enhancement of the system.

- *Motivation*

The main motivation behind PictoVoice is the need for a simple way to make visual information easier to access through audio. An image can contain several objects, people, and activities, but understanding all of this information directly from the image is not always convenient. Providing a spoken description can offer an alternative way of accessing the same information.

Existing image captioning techniques can generate textual descriptions from images, while speech synthesis can convert text into audio. Bringing these two capabilities together makes it possible to move from visual content to a spoken description through a single workflow. This motivated the development of PictoVoice as a system that combines image processing, visual feature extraction, caption generation, and speech output.

Another motivation for the project is accessibility. A system that can describe an image and read the description aloud can be useful in situations where visual information needs to be accessed through audio. The project also provides scope for supporting different languages for multilingual support. This multilingual functionality is currently under implementation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep311

➢ *Existing System*

Several methods are already available for helping computers understand the content of images. Image classification methods mainly identify objects or categories in an image, whereas image captioning systems attempt to describe the image using natural language. With the use of deep learning, these systems have become better at extracting useful information from visual content.

Many image captioning approaches use one model to understand the image and another model to generate the description. CNN-based networks can be used to obtain visual features, which are then provided to language models such as LSTM for generating a caption. Some systems have also extended image captioning by adding speech output so that the generated information can be listened to rather than only read.

However, these functions are not always available together in a single simple system. A captioning application may provide only text, while a speech-based application may require a separate method for obtaining the description from an image. This can make the overall process less convenient for users who need audio-based access to visual information.

PictoVoice brings these stages together in one workflow. The uploaded image is processed to obtain visual features, a caption is generated from those features, and the resulting caption is converted into speech. This provides a direct connection between the original image and its spoken description.

II. LITERATURE SURVEY

Research on automatic image description has gradually moved from simple visual recognition toward systems that can produce language that is understandable to users. This area brings together visual processing and language generation, and it has also been explored as a way of making image-based information easier to access through audio [1], [4].

A common approach is to first obtain a compact representation of the image and then use a sequence model to produce a description. CNN models are suitable for learning visual representations, while LSTM networks can use those representations together with previously generated words to construct a caption step by step. CNN–LSTM based systems have therefore been widely explored for automatic image description [7]. Voice-assisted approaches extend this idea by providing the generated information through speech rather than depending only on a visual or textual response [2], [6].

Several studies have also investigated attention mechanisms for image captioning. Instead of relying only on a single representation of the complete image, attention allows the captioning model to give greater importance to relevant visual regions during different stages of sentence generation. Different CNN architectures have been evaluated along with such approaches to improve the representation used by the captioning model [5].

Another important research direction is the conversion of generated captions into audio. In these systems, image understanding and language generation are followed by a speech synthesis stage. This combination is particularly useful for accessibility applications because the information generated from an image can be consumed as spoken output. Previous work has investigated such image-to-speech and audiodescription systems for users who may have difficulty accessing visual information directly [1], [3], [4].

More recent work has examined Transformer-based architectures for image captioning and vision-to-voice applications. These models offer an alternative to recurrent sequence generation and are being investigated for their ability to capture relationships within visual and textual information [8], [9]. Research has also considered caption generation in languages with fewer available resources, which is relevant when extending accessibility systems beyond English [10].

The reviewed studies address different parts of the overall problem, including visual representation, sequential caption generation, attention-based processing, speech output, and newer Transformer architectures [1]–[10]. These works provide useful guidance for designing a system in which visual analysis, caption generation, and speech synthesis operate as connected stages.

PictoVoice applies these ideas in a single workflow. The implemented system accepts an image, prepares it for processing, extracts visual features using ResNet-50, generates a description using an LSTM-based model, and converts the resulting text into speech. Compared with treating caption generation and speech synthesis as separate tasks, this arrangement provides a direct path from the uploaded image to an audible description. Multilingual support is being developed as an additional enhancement to the current system.

III. PROPOSED METHODOLOGY

PictoVoice follows an end-to-end workflow in which an input image is processed, analyzed, described in text, and finally converted into speech. The methodology combines image preprocessing, deep visual feature extraction, caption generation, and text-to-speech conversion. Each stage produces an output that is used by the following stage, allowing the complete process to be carried out as a single workflow.

➢ *Image Input and Preprocessing*

The process begins when the user selects and uploads an image through the PictoVoice interface. The input image is prepared before it is given to the deep learning model. The image is resized and normalized to match the requirements of the feature extraction stage. This preprocessing step provides a consistent representation of the input and prepares the image for further analysis.

➢ *Visual Feature Extraction*

After preprocessing, the image is passed to the visual feature extraction stage. ResNet-50 is used as the deep convolutional neural network for extracting important visual

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep311

information from the image. Instead of directly classifying the image, the network is used to obtain a feature representation that captures relevant information about the visual content. This representation is then forwarded to the caption generation stage.

➢ *Caption Generation*

The extracted visual representation is provided to an LSTMbased caption generation model. The LSTM processes the visual information along with the previously generated words and predicts the next word in the sequence. This process continues until a complete description of the image is generated. The resulting caption provides a textual representation of the visual content identified by the system.

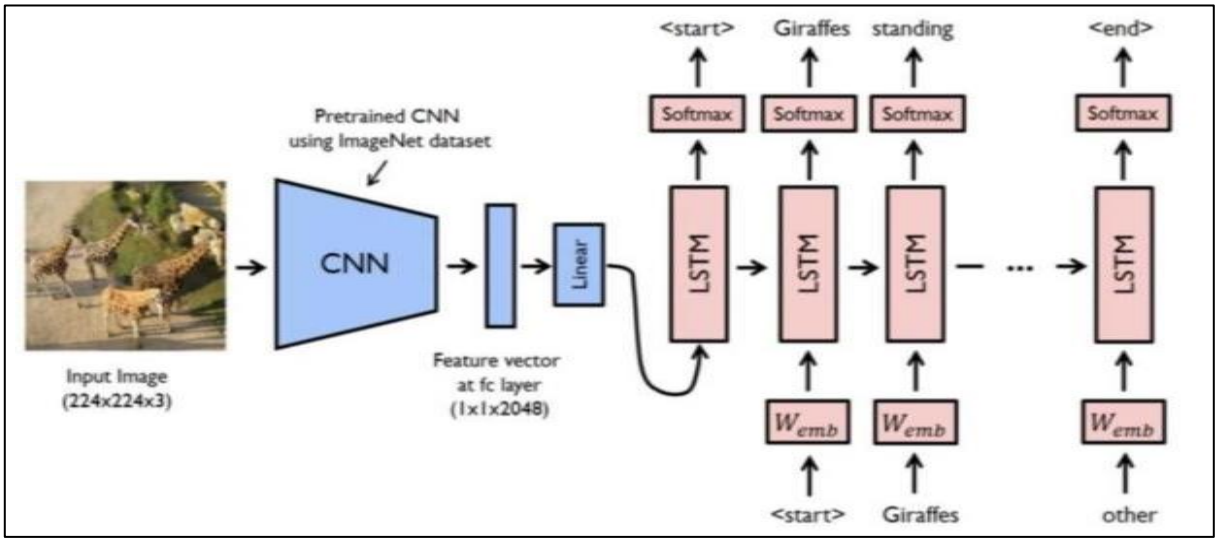

Fig 1 CNN–LSTM Based Image Caption Generation Architecture.

➢ *Speech Generation*

Once the textual caption has been generated, it is passed to the Text-to-Speech module. The speech component converts the caption into spoken audio, allowing the user to listen to the description instead of depending only on the displayed text. This stage completes the image-to-speech workflow of PictoVoice.

➢ *User Interaction*

The system is designed to keep the interaction simple for the user. The user uploads an image through the interface, waits for the system to generate its description, and can then use the available speech output option to listen to the generated caption. This provides a direct interaction path from image input to spoken information.

➢ *Multilingual Support*

Multilingual support is being developed as an extension of the current system.This functionality is currently under implementation and is therefore not considered part of the completed image-to-speech functionality.

➢ *System Architecture*

The system architecture of Picto Voice illustrates the process of generating a textual description from an input image using a CNN–LSTM based image captioning approach. The input image is first processed by a convolutional neural network to extract relevant visual features. The extracted feature vector is then provided to the LSTM network, which generates the image caption sequentially. The output words are predicted using the softmax layer at each time step.

IV. IMPLEMENTATION

The PictoVoice system is implemented as a sequence of connected modules, where the output from one stage becomes the input for the next. The implementation focuses on taking an image uploaded by the user, extracting its visual information, generating a textual description, and converting that description into speech. This provides a complete workflow from image input to audio output.

➢ *Image Processing*

The implementation begins with image acquisition through the user interface. The selected image is prepared before it is passed to the deep learning model. The image is resized and normalized so that it matches the expected input format of the feature extraction network. This step provides a consistent input representation for the subsequent processing stages.

➢ *Feature Extraction*

ResNet-50 is used to extract the visual representation of the input image. The network processes the preprocessed image through its convolutional layers and produces a feature representation containing important information about the visual content. These extracted features are not used simply to assign an image class; instead, they serve as the visual representation required by the caption generation stage.

➢ *Caption Generation*

The visual features obtained from ResNet-50 are passed to the LSTM-based caption generation module. The LSTM processes the visual representation and generates the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep311

description sequentially. It predicts the next word based on the available visual information and the previously generated words. The process continues until the description is completed, producing a textual caption that represents the main content of the input image.

➢ *Text-to-Speech Conversion*

The generated caption is then provided to the Text-toSpeech module. The TTS component converts the textual description into spoken audio. The generated speech can be played through the interface using the available listening option. This allows the final information to be accessed as audio rather than only as text.

➢ *User Interface and Output*

The user interface provides a simple way to interact with the different stages of the system. The user can select an image, submit it for processing, view the generated caption, and access the speech output. The system therefore produces both a textual description and an audio representation of the generated caption.

The implementation demonstrates the integration of computer vision, language generation, and speech synthesis in one application. The current implementation completes the imageto-speech functionality, while multilingual narration is being developed as an additional feature.

Fig 2 Picto Voice Image Upload Interface.

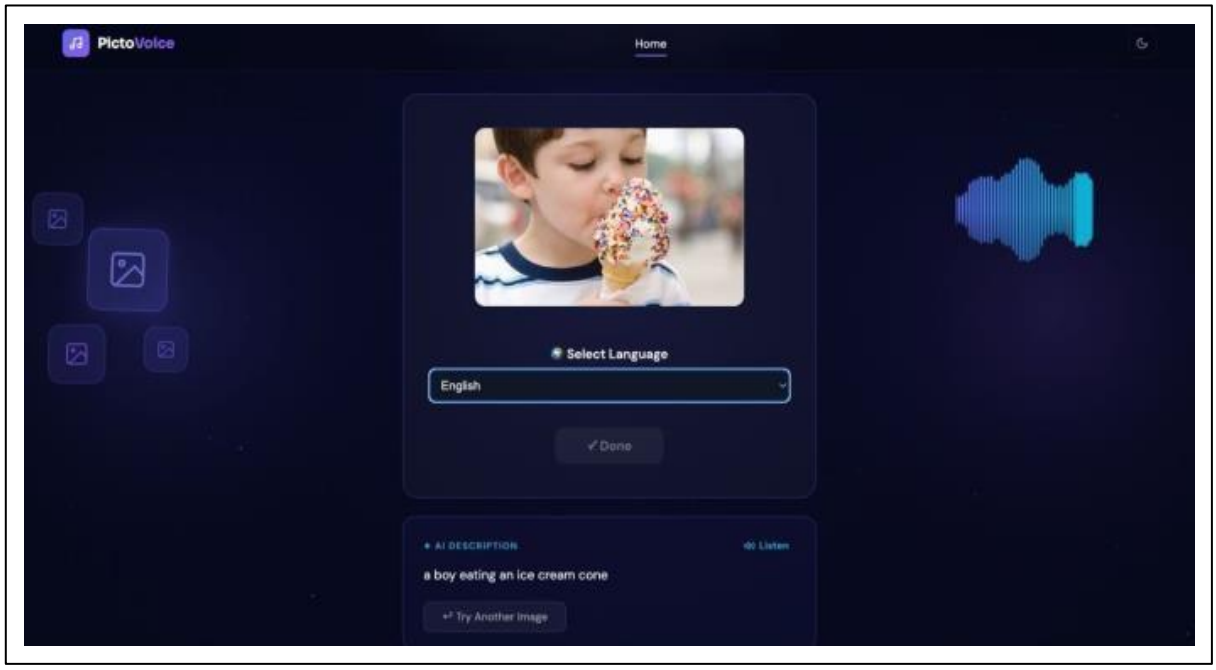

Fig 3 Generated Image Description and Speech Output Option.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep311

V. RESULTS AND DISCUSSION

The implemented Picto Voice system was tested to evaluate the complete image-to-speech workflow. The system accepts an image as input, processes its visual content, generates a textual description, and provides the generated description as speech output.

Figure 2 shows the main Picto Voice interface. The interface provides an image upload option through which the user can select an image for processing. The simple upload-based interaction allows the user to provide visual content to the system.

Figure 3 shows the result produced by the system for the selected input image. The system generates the description "a boy eating an ice cream cone" based on the visual content. The interface also provides a "Listen" option, through which the generated description can be converted into speech and presented to the user.

The results demonstrate the integration of image processing, visual feature extraction, image caption generation, and speech generation into a unified workflow. The system successfully produces a meaningful textual description and provides an audio-based way of accessing the generated information.

The current implementation demonstrates the image-tospeech functionality. Multilingual support is currently under implementation and will be evaluated as part of the future enhancement of the system.

VI. ADVANTAGES AND APPLICATIONS

PictoVoice combines image understanding and speech generation in a single workflow, providing several advantages for users who need to access visual information through audio.

➢ *Advantages*

One of the main advantages of the system is its simple interaction model. The user can upload an image, obtain a generated description, and listen to the corresponding speech output without having to manually describe the image. The integration of image processing, caption generation, and speech synthesis also reduces the need to use separate tools for these tasks.

The system provides both textual and audio forms of the generated information. This makes the output more flexible and can be useful in situations where listening to a description is more convenient than reading or interpreting an image directly. The modular structure of the system also allows individual components to be improved or extended without changing the complete workflow.

Another advantage is the possibility of extending the system with additional languages and improved captioning techniques. Multilingual support, attention-based models, and transformerbased approaches can be incorporated as the system develops further.

➢ *Applications*

- Accessibility: PictoVoice can be used as an assistive technology to provide spoken descriptions of images, helping users access visual information through audio.
- Education: The system can be used with educational images, illustrations, and other visual learning materials where an audio description can provide an additional way of accessing the content.
- Digital Content: The captioning capability can be useful for organizing and describing images in digital content and multimedia applications.
- Smart Assistive Technologies: The current system can serve as a foundation for future assistive applications in which images captured by a camera are analyzed and described through speech.
- Human–Computer Interaction: By allowing a system to interpret visual information and communicate the result through natural language and speech, PictoVoice can contribute to more accessible forms of interaction between users and intelligent systems.

VII. LIMITATIONS

Although PictoVoice successfully demonstrates the complete workflow from image input to speech output, the current system has some limitations. The quality of the final description depends on how well the caption generation model interprets the visual information in the input image. Images containing several objects, activities, or complex relationships may therefore result in descriptions that do not capture every detail of the scene.

The current implementation focuses on generating a single description for an input image. It does not yet provide detailed scene-level analysis or multiple descriptions for different regions of an image. Improving the ability to understand complex scenes can make the generated captions more informative in future versions.

Another limitation is that multilingual support is still under implementation.

The quality of the speech output is also dependent on the generated caption. If the caption does not accurately represent the visual content, the corresponding audio description will have the same limitation. These aspects provide opportunities for improving the captioning model and extending the system in future work.

VIII. CONCLUSION

PictoVoice demonstrates a practical approach for converting visual information into spoken descriptions. The developed system brings together image processing, visual feature extraction, image caption generation, and Text-to-Speech conversion to form a complete workflow from an uploaded image to an audio description.

The implemented system uses ResNet-50 to obtain a visual representation of the input image and an LSTM-based

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26sep311

model to generate a textual description from the extracted features. The generated caption is then converted into speech, allowing the user to access the description through audio. The developed interface also provides a simple interaction through which an image can be uploaded, its description can be viewed, and the corresponding speech output can be accessed.

The current implementation demonstrates the basic imageto-speech functionality of PictoVoice and provides a foundation for further development. Multilingual support is currently under implementation, while improvements in scene understanding, caption generation, and speech output can further enhance the system. Overall, the project shows how computer vision, natural language generation, and speech synthesis can be combined to make visual information more accessible.

FUTURE ENHANCEMENT

The current version of PictoVoice establishes the basic image-to-speech workflow, but several improvements can be explored to make the system more capable and useful in practical situations.

One of the immediate enhancements is the completion of multilingual support so that users can receive image descriptions in a language that is more familiar to them. This can make the system more useful across different linguistic backgrounds.

The caption generation component can also be improved by introducing attention-based techniques. An attention mechanism can help the model focus on relevant regions of an image while generating different parts of the caption. Transformerbased models can also be explored to improve contextual understanding and produce more detailed descriptions.

Another possible extension is real-time image and video description. Instead of processing only an uploaded image, the system could eventually analyze frames from a camera and provide spoken descriptions of the surrounding content. This could make the system more suitable for real-world assistive applications.

The system can also be extended to mobile and cloudbased platforms. A mobile version could allow users to capture an image directly using a device camera and receive its spoken description. Cloud deployment could provide access to more computationally intensive models and support largerscale applications.

These enhancements can further improve the accuracy, accessibility, and practical usefulness of PictoVoice while providing a foundation for future research in intelligent image understanding and assistive technologies.

REFERENCES

[1]. S. Shanthi, P. Gowthami, T. S. Harshitha, T. Kavya, and K. Sangeetha, “Deep Learning based Audio Description of Visual Content by Enhanc-ing Accessibility for the Visually Impaired,” in Proc. IEEE Int. Conf. Sustainable Communication Networks and Application, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10370414/

[2]. “Computer Vision and Voice Assisted Image Captioning Framework for Visually Impaired Individuals using Deep Learning Approach,” IEEE, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/ 10353449/

[3]. “An Efficient Image to Speech Generation System for Visually Impaired Individuals Using Deep Learning Techniques,” IEEE, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10717315/

[4]. “Machine Learning based approach to Image Description for the Visu-ally Impaired,” IEEE, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9544867/

[5]. “Comparative Analysis between InceptionResnetV2 and InceptionV3 for Attention based Image Captioning,” IEEE, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9587514/

[6]. “Deep Learning Based Voice Assistant for the Visually Impaired,” IEEE, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/ 9532661/

[7]. “Deep Fusion: A CNN-LSTM Image Caption Generator for En-hanced Visual Understanding,” IEEE, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10351389/

[8]. “Deep Learning for Automated Image Captioning: A CNN and Trans-former Model Analysis,” IEEE, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10914740/

[9]. “Transforming Disability Into Ability: An Explainable Vision-to-Voice Image Captioning Framework Using Transformer Models and Edge Computing,” IEEE Journal, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/11195080/

[10]. “Image Captioning for the Visually Impaired and Blind: A Recipe for Low-Resource Languages,” IEEE, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10340575/