International Journal of Research in Engineering, Science and Management
Volume 5, Issue 6, June 2022
https://www.ijresm.com | ISSN (Online): 2581-5792

Audio Narration of a Scene for Visually Disabled using Smart Goggle: An Implementation Survey

Pratyush Pratap Singh[1], Sharath S. Hegde[2], R. Varun[3*], Vivek Hegde[4], K. A. Sumithra Devi[5]
[1,2,3,4]*Student, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, India*
[5]*Head of the Department, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, India*

***Abstract*: Blindness is one of the major problems in our society which made it difficult t for a blind person to lead their day-to-day life. Due to the loss of sight, many information about the real-world scenarios and surroundings will be missed for blind people. By using various multimedia information processing methods and technologies, our device will first acquire the image information through either Pi camera or Webcam, then performs and image to text conversion. Then the blind person is directed through voice commands generated by the application according to the object detection. Before only text reading system was there. But in this approach our Smart Goggle system will detect, track and analyze the approaching objects and give information about them. This will help the blind people to get more information access about real world and reduces numerous difficulties faced by them.**

***Keywords*: Raspberry Pi, TensorFlow, TensorFlow Lite, OpenCV, Convolution Neural Network, Text to Speech (TTS) engine, Object detection, Long Short-Term Memory (LSTM), Recurring Neural Network.**

1. Introduction

For people with normal vision, more than 80% of the external information is acquired through visual channels only. According to World Health Organization, in 2011, there is an estimation of more than 2 billion visually disabled people in the whole world, in which an approximation of 13% are fully blind and rest are with low vision. The old mobility aid like walking cane has its own drawbacks with respect to range of motion and amount of information.

In 2018, a project called "Tech for Social Good" was officially launched by Tencent Research Institute. “Human is the scale of technology” was the main aspect of this project. Inspired by this idea, we aim to design and develop a wearable audio-visual assistive device for the visually impaired to enable them to perceive the real-world information, in an attempt to make the world better for the blind community. There are many solutions existing for the problem of assisting people with vision impairment. “Automatic Image Annotation “[1] is helpful in several applications like image indexing, image retrieval and increasing accessibility to users. But a list of labels sometimes becomes indefinite. Hence, there is an urgent need for automatic conversion of image annotations to natural language, which will have a stronger semantic content. This device will help the blind people to perceive the outside world, reduces the numerous difficulties faced by them in their day-to-day activities like navigation, obstacle detection, locating objects, etc. Also provides the visually impaired person a greater level of independence.

2. Experimental Survey

Pi camera or Webcam is directly connected to Raspberry pi. Captured images from the surrounding are fed into the Raspberry Pi [2], [3], [5], [6], [8], [9]. V. V. Mainkar (2020) [8] used a software called Imagemagick to enhance the captured image to smoothen it in the processing stage. L. George (2020) [2] and Rithika H (2016) [9] used Tesseract Library to get the text extracted from images. S.Thiyagarajan (2018) [6], G.Sekar (2021) [5] and Thomas, A. (2020) [2] used OpenCV Library to detect the objects from captured images. Although above research works used OpenCV Library which gave only labels for the detected objects, however this approach does not generate human understandable form of description or sentences. So, A. Gupta (2012) [1] used Automatic Image Description generating techniques to resolve this issue.

3. Methodology

Once the system is turned on, there will be an alert signal which says 'Start'. Then, the user can either choose to detect the object or get the image description. Once done, the system can halt using 'Stop' button. There are mainly two functionalities, Object Detection and Scene Narration.

A. Object Detection

In first process, Real-time Capturing is the first step. This is done using either Pi camera or Webcam. When we annotate an image in Real-time, we are adding metadata to a dataset. Each image in dataset must be thoughtfully and accurately labeled to train an AI system to recognize objects similar to the way a human can. Image labeling is the process of identifying and marking various details in an image. Extraction of such details into a digital text is required in label extraction. Text-to-speech (TTS) reads digital text aloud. This section performs the task of converting the machine-coded text content into an audio signal format. The result of the Text to Speech module will be sent to

*Corresponding author: varun.r.mrv@gmail.com

the Bluetooth headset (which is already in pair with Raspberry Pi), and the audio transmission signal is taken as the output. In this way, text as an audio voice is easily heard and understood by blind people.

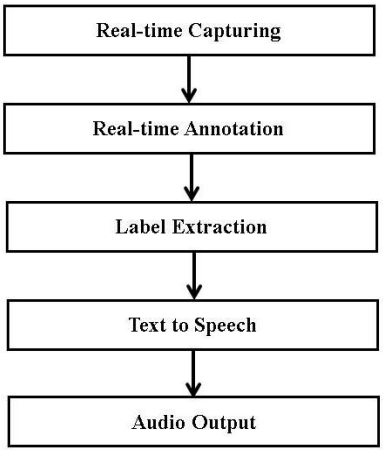

Fig. 1. Flow of Process in process 1

B. *Scene Narration*

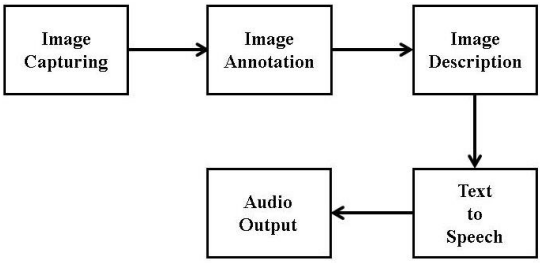

Fig. 2. Flow of Process in process 2

The second process is also similar to the first process. In this process, we assign labels to captured image and get the description based on annotation in the form of sentence. Then that text sequence is converted to audio using pyttsx3, which is a text-to-speech conversion library in Python. The output of Text to speech module will be sent to Bluetooth device paired with Raspberry pi and audio signal can be heard by blind people.

4. Implementation

Some of the tools and techniques used in development of this system design are given below.

A. *Build model*

It contains the files for constructing, training and evaluating the deep-learning model.

B. *Caption generator*

It contains the code to use the built model to generate captions for the images in sample images.

C. *Convolutional Neural Networks*

CNN is a subfield of Deep learning and specialized deep neural networks used for the recognition and classification of images. It is used to process the data represented as a 2D matrix like images.

D. *Long short-term memory (LSTM)*

Being a type of RNN (recurrent neural network), LSTM (Long short-term memory) is capable of working with sequence prediction problems. It is mostly used for the next word prediction purposes, as in Google search our system is showing the next word based on the previous text.

E. *TensorFlow*

TensorFlow is an end-to-end open-source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications.

F. *TensorFlow Lite*

TensorFlow Lite is a mobile library for deploying models on mobile, micro controllers and other edge devices.

G. *OpenCV*

OpenCV is the huge open-source library for the computer vision, machine learning, and image processing and now it plays a major role in real-time operation which is very important in today's systems. By using it, one can process images and videos to identify objects, faces, or even handwriting of a human. To Identify image pattern and its various features we use vector space and perform mathematical operations on these features.

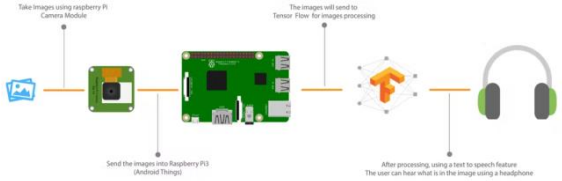

Fig. 3. Shows a simplified sketch of the design concept

5. System Architecture

A. *Hardware Requirements*

The hardware parts of the system include:

- Raspberry Pi 4 (8GB Ram)
- Webcam or Pi Camera
- Press Buttons
- Micro SD Card (64 GB)
- Power Bank Unit
- Bluetooth Headset
- Goggle
- Connecting Cables

B. *Datasets Needed*

The Flickr30k dataset has become a standard benchmark for sentence-based image description. This paper presents Flickr30k Entities, which augments the 158k captions from Flickr30k with 244k coreference chains, linking mentions of the same entities across different captions for the same image, and associating them with 276k manually annotated bounding boxes. Such annotations are essential for continued progress in automatic image description and grounded language understanding.

Imagenet dataset is used to train the CNN model called Xception. Xception is responsible for image feature extraction.

6. Conclusion

We have come to the conclusion that our voice assisted scene narration based Smart Goggle will be useful for people with vision impairment. The proposed combination of design is cheaper and an efficient assistive device for the blind that generates a real-time, high-quality image descriptions by detecting the object in front of the user and provide the alert feedback, thus making navigation more safe and secure. If a scene is detected within the field of view, the image of the scene will be extracted and converted into text sequence. Finally, the text sequence will be fed back to the user via voice output. Since the output of the device is audio, blind people can easily hear it, making it a very useful device for the blind.

References

[1] Gupta, Ankush, and Prashanth Mannem. "From image annotation to image description." International conference on neural information processing. Springer, Berlin, Heidelberg, 2012.

[2] George, L., Rinsila, S., Baby, R., & Thomas, A. (2020). Raspberry Pi based Reader for Blind. J. Opt. Commun. Electron., 5(03), 11-16.

[3] J. Ai et al., "Wearable Visually Assistive Device for Blind People to Appreciate Real-world Scene and Screen Image," 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), 2020, pp. 258-258.

[4] Lee, Seung-Jun & Lee, Yong-Hwan & Ahn, Hyochang & Rhee, Sang-Burm. (2021). Color Image Descriptor using Wavelet Correlogram.

[5] M. I. S, G. R. R, D. R and G. Sekar, "Smart Obstacle Recognition System using Raspberry Pi," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 672-675.

[6] Kumar, G. & e, Praveen & Sakana, G., (2018). Implementation of Optical Character Recognition Using Raspberry Pi for Visually Challenged Person. International Journal of Engineering and Technology (UAE).

[7] Vadwala, Ayushi Karmakar, Yesha Suthar, Krina Thakkar, Nirali. (2018). Object Detection System using Arduino and Android Application for Visually Impaired People. International Journal of Computer Applications. 181. 975-8887.

[8] V. V. Mainkar, T. U. Bagayatkar, S. K. Shetye, H. R. Tamhankar, R. G. Jadhav and R. S. Tendolkar, "Raspberry pi based Intelligent Reader for Visually Impaired Persons," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2020, pp. 323-326.

[9] H. Rithika and B. N. Santhoshi, "Image text to speech conversion in the desired language by translating with Raspberry Pi," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016, pp. 1-4.