Audio Narration of a Scene for Visually Disabled using Smart Goggle

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]*Professor, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, India*

***Abstract*: This work supports visually disabled people to get an idea of what is in the captured image. By using different kinds of multimedia information processing techniques, the proposed device will first acquire image attributes via Pi Camera, then perform an image to text conversion using Tesseract library and OpenCV library. Previously proposed approaches used computer vision technology to determine labels or exploit already available descriptions of the training images to transfer or compose a completely new description for the image to be tested. Now we propose an approach that will use image annotations to generate image descriptions and shows that with the accurate object and attribute detection, human-like descriptions for images can be generated. We use TTS (Text to Speech) for text to speech transformation and Python programming language.**

***Keywords*: Raspberry Pi, Tesseract OCR engine, Raspberry Pi camera board, OpenCV, Natural Language Processing, Natural Language Generation, Text to Speech (TTS) engine, Optical Character Recognition (OCR), Object detection.**

1. Introduction

In 2018, Tencent Research Institute officially launched a project called "Tech for Social Good". The main aspect of this project was that “human is the scale of technology”. Motivated by this idea, we aim to develop a visually assistive device for the blind people to enable them to perceive the outside world, in an attempt to make it better for the blind community.

There are numerous existing solutions to the problem of assisting people who are visually disabled. “Automatic image annotation “[1] is beneficial in several applications like image retrieval, image indexing and increasing accessibility to users. But a list of labels sometimes becomes indefinite. For example, an image annotated with labels {green, airport, jet} does not communicate the information whether attribute green is associated with airport or jet and whether “jet is taking off from airport” or “jet is parked at the airport”. We focus on improving the competence of blind people by providing them with a solution where the details are given in the form of audio signal.

2. Related Work

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

3. Proposed Approach

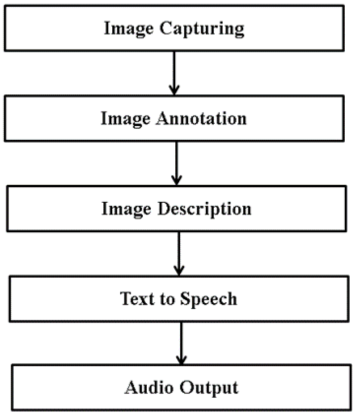


Fig. 1. Flow of process

A. *Image Capturing*

Image acquisition is the first step. This is done using a push button that connects to the Pi. High-resolution cameras are used to improve image quality.

B. *Image Annotations*

Here we train the model using some dataset so that all objects in the image are labeled. Within the snapshot, some objects exist. All objects are detected by employing some libraries or collections but all objects are not examined, while the object which has more precision can be examined and analyzed for results.

C. *Image Annotations to Description*

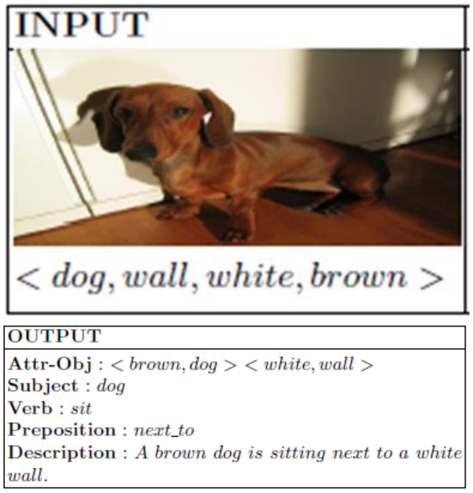
A brief description of the image "A White dog is lying in bed" contains below mentioned information:

i. Objects present in the image: <dog>, <bed>

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

ii. Attribute: <white>
iii. Attribute-Object pair: <white; dog>
iv. Subject: <dog>
v. Verb: <lie>
vi. Preposition: <on>

Starting with the image and its annotations ((a), (b)), our task is to create the description of the image. The majority of the prior methods used the "PASCAL1" sentence dataset, so we will again use the same to report the results. “PASCAL” dataset consists of 1000 images, each image containing 5 human-generated descriptions that give the image insight.



Attr-Obj : < *brown, dog* > < *white, wall* >
Subject : *dog*
Verb : *sit*
Preposition : *next_to*
Description : *A brown dog is sitting next to a white wall.*

Fig. 2. Overview showing the usage of annotated image to automatically generate description

We map each "attribute" to the corresponding "object", determine the "subject" of the sentence, predict the "verb", "preposition", "determiner" from the available information of the training image, to acquire four pairs ((det1, attribute1, subject), verb, preposition, (det2, attribute2, object)) which are utilized to generate the components of the image. These pairs are depicted with an appropriate image in the above figure 2.

D. *Text to Speech*

This module performs the task of converting the machine-coded textual content into an audible format. It is here, we represented a system to scan written text, for helping the blind individuals. Word recognition on the text regions is performed using OCR. Text-to-Speech synthesis is used to pronounce the document through the ear phones.

It is a software program used to synthesize speech from textual content [3]. The TTS engine converts the written text into a phonemic representation, which translates into a waveform that can produce a valid speech output. eSpeak [8] is a software program that can be utilized on Raspberry Pi by installing the eSpeak engine effortlessly. Here it is used to modify and convert a text file to an audio file with a .flac extension.

E. *Audio Output*

The result of the speech engine/synthesizer is enhanced by an audio amplifier, then it is sent to the Bluetooth headset (already paired with the Raspberry Pi), and the sound transmission signal is taken as the output. In this way, text as the tone of voice is easily heard and understood by people with visual impairments.

4. Architecture of the Proposed System

A. *Hardware Requirements*

The hardware parts of the system include:

- Raspberry Pi 4 (8GB Ram)
- Pi Camera module(8MP)
- Press Button
- Memory Card (32 GB)
- Power Bank
- Bluetooth Earphone
- Goggle
- Connecting Cables

B. *Datasets Needed*

In our approach we use the “UIUC PASCAL Sentence” dataset and the “IAPR TC-12” Benchmark to test the efficiency and performance of our approach. “PASCAL” dataset contains 1000 images each having 5 human-written descriptions. The “IAPR TC-12” benchmark comprises of 20,000 images each containing a description of up to 5 sentences. Unlike “PASCAL” dataset, due to the number of complex images and complex descriptions in the dataset, it makes automatic generation of image descriptions more challenging.

5. Conclusion

After thorough survey we have come to the conclusion that voice assisted scene narrating system based Smart Goggle will be useful for visually disabled people. The results of experiments conducted by various researchers showed us promising results for the images that are captured. This paper presented a novel approach to generating real-time, high-quality image descriptions from an annotated image. Their methodology processed the captured image and read it out clearly. Since the output of the device is voice, visually impaired people can easily hear it, making it a very useful device as well as economically advantageous 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).7.65-67.

[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.