iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35244

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

Caption to Voice Bot for Assistive Vision

Nandita S[1], Raashi A S[2], Prajwala A N[3], Vinod Kumar H[4]
[1, 2, 3, 4]*Department of Telecommunication Engineering, Dayananda Sagar College of Engineering*

Abstract: ***Over the last few years, with the rapid development of artificial intelligence, the generation of the caption of images has progressively caught the considerable interest of several artificial intelligence research groups and has become a fascinating and tedious mission. A large component of scene comprehension, which encompasses the knowledge of computer vision and natural language processing, is image caption, which automatically produces natural language explanations according to the content observed in an image. The applications of such an image caption are substantial and noteworthy. The prime intention of the project is to build an object detection and captioning module that produces captions from the features extracted from the input images fed to the module in the form of audio and interface it with a virtual text reader, a read-aloud technology. Additionally, both these features can be accomplished using live images. The module as a whole helps the visually impaired identify objects and their positions.***

Keywords: ***Convolutional Neural Networks, Tesseract OCR, GTTS, Object Detection, Caption Generation***

I. INTRODUCTION

Having the ability to see is a wonderful gift. Individuals with vision are able to see and interpret the environment around them. Visual deficit is a state in which one is unable to recognise objects due to physiologic or neurological factors. Visual impairment can make it difficult for persons to carry out day-to-day tasks. As per a recent estimate, 253 million people worldwide suffer from visual loss. There are 36 billion people who are blind, and 217 million people who have moderate to severe vision problems [14]. Individuals and their family suffer tremendously as a result of their loss of sight. This system is designed to help visually impaired persons live a more independent life. It is integrated with cutting-edge technology and is designed to let the visually impaired lead a life free of limitations. This is a visual-based application with a few major components such as a camera, a system containing OpenCV, and speakers connected together, as well as additional internet-based working techniques. The project's input is an image/video (several frames), which will be captured and analysed using a camera connected to the computer. As a result, the object gets identified, and audible data is transmitted to the blind individual through speakers or headphones.

II. LITERATURE SURVEY

A. Small-Object Detection Based on Deep Learning: Wei Wei [1]

This research introduces a deep learning-based object detection algorithm. This study primarily discusses the detection algorithm based on regional suggestion and regression and examines the detection method's merits and weaknesses and detection performance in terms of accuracy and speed. The shortcomings of various detection approaches in detecting small items and the challenges in detecting small items are then discussed. Firstly, the public data sets and evaluation criteria for small object detection are introduced on this foundation. Object detection methods based on classification are also known as two-stage algorithms since they are divided into two steps. The algorithm should first extract the candidate region, after which it should determine the candidate area's category and change its placement. Finally, output the object detection result. R-CNN, SPP-Net, Fast R-CNN, and Faster R-CNN are examples of R-CNN series algorithms based on region extraction. Second, the one-stage approach is a regression-based object detection system that simplifies the object detection process by combining end-to-end regression problems. Unlike the two-stage approach based on region extraction, the one-stage approach can achieve feature sharing in a single training and has substantially increased detection speed. YOLO, SSD, and other one-stage object identification methods are common examples. Third, small object detection can be used in unmanned aerial vehicle ground object identification, pedestrian detection in traffic scenes, traffic sign detection, and other situations.

B. Image Caption Generation using a Deep Architecture: Ansar Hani et al [2]

This paper presents a model which describes an image in Natural Language using Computer Vision (CV), Natural Language Processing (NLP), and Machine Learning methods. The three measures for captioning include CNNs, RNNs, and the Attention Mechanism. Convolutional Neural Networks are being deployed as encoders to derive an image's features and then these are fed into Recurrent Neural Networks for Language Modelling. Additionally, an Attention Mechanism is incorporated while producing captions. CNN's as an encoder constructs a rich image representation by embedding it in a fixed-length vector representation.

©IJRASET: All Rights are Reserved

An Inception V3 Model is pre-trained with ImageNet/MSCOCO dataset, this paper uses MS-COCO dataset as it has instance-level segmentation followed by Image feature extraction. TensorFlow 2.0 is used to implement the Model. Owing to the representations of the image, a decoder is used to transform the image into natural sentences. RNNs as decoders are implemented using LSTM or GRU.

C. Deep Learning Based Automatic Image Caption Generation: Varsha Kesavan et al [3]

This system uses a transfer learning approach to generate automatic captions for any given image. The database of images is provided as an input to the deep neural network, i.e. The CNN encoder is used to create a "thought vector" that derives the image characteristics, and the RNN decoder is used to translate the image characteristics and artifacts to obtain a systematic and relevant image definition. The VGG16 pre-trained model is the encoder included in this model. This system uses a recurrent neural network that encodes the variable length input into a fixed dimensional vector and decodes the required output phrase using this definition. The vector containing the output of the fully connected layer in VGG16 is named the thought vector, which is linked to GRU units. The final output is the description of the image in plain English.

D. An Intelligent Approach of Text to Speech Synthesizers for English and Sinhala Languages: Pabasara Jayawardhana et al [4]

This paper aims to investigate the new Deep Voice-based Text-to-Speech (TTS) algorithm, which is a completely convolutionary mechanism based on attention. WaveNet, Merlin, and Deep Voice are among the many TTS versions that have been released. This study uses the most recent Deep Voice version, Deep Voice 3. There are three components of the Architecture of Deep Voice 3, i.e., Encoder, converter, and decoder. Encoder: It is a deep convolutional neural network that transforms textual characteristics to an internal representation that has been studied. Decoder: The decoder is used to decode a low-dimensional audio representation (or Mel spectrograms) for the learned representation coming from the encoder. It consists of casual convolutions with convolutionary multi-hop attention and autoregressive generated output. Converter: It's a completely-convolutionary post-processing network. The converter forecasts final parameters from the hidden states of the decoder. The conclusions of the experiments specify that text-to-speech neural network-based systems have the potential to transmit better voice quality than conventional approaches, although some system improvement is still required.

III.METHODOLOGY

Image captioning, which automatically provides natural language explanations based on the content shown in an image, and other aspects that help develop an assistive environment for the visually impaired are discussed here. In our suggested architecture, other characteristics that make up a protected environment for the blind include a virtual text reader that works in tandem with object recognition and captioning modules. Object detection is a Computer Vision challenge that involves identifying and detecting objects of specific classes in an image and then creating captions that are then turned to voice to autonomously navigate visually impaired people and let them live a normal, independent life. Object localization can be done in a variety of ways, such as drawing a bounding box around the object or identifying every pixel in the image that contains the object (a process known as segmentation), and the captions can be converted to speech using the Google Text to Speech python package. The use of virtual text reader technology allows the visually handicapped to comprehend the content of newspapers, books, magazines, journals, and other publications, which can be used for recreational or educational purposes. The virtual text reader can also be done with live visuals, allowing the visually handicapped to read names of packages, medicines, and other items that are right in front of them.

Packages and libraries used are Tesseract OCR, OpenCV, TensorFlow, Keras, NumPy, gTTS, Playsound, Imutils, Argparse and Pillow.

A. Object Detection and Captioning

The Object detection is a computer vision approach for identifying and locating things in images and videos. Object detection, in particular, creates bounding boxes around identified items, allowing them to see where they are and how they move through in a scene.

1) *Dataset:* The MS COCO Dataset which stands for Microsoft Common Object in Context, is intended to represent a wide range of objects that we come across on a daily basis. The COCO dataset is labelled, providing information for training supervised computer vision models that can recognize the dataset's common objects. Of course, these models are far from flawless, thus the COCO dataset serves as a baseline for evaluating the models' progress over time as a result of computer vision research. It also serves as a training dataset for computer vision models. The model can be fine-tuned to learn various tasks with a custom dataset after it has been trained on the COCO dataset [15]. The COCO dataset includes the following items:

©IJRASET: All Rights are Reserved

a) There are 121,408 photos in the COCO Dataset.
b) There are 883,331 object annotations in the COCO Dataset.
c) There are 80 classes in the COCO Dataset.
d) The median picture ratio in the COCO Dataset is 640 × 480 pixels [15].

2) Algorithm: You Look Only Once (YOLO) is a sophisticated real-time object detection convolutional neural network (CNN). Object detection is performed on an unseen snapshot using a pre-trained model. This is included in a single Python file with approximately 435 lines of code. This script is a software that will prepare a model using pre-trained weights, then use that model to do object detection and produce a model. It is also reliant on OpenCV. Rather than utilising this software directly, reuse parts of it and write own scripts to create a Keras YOLOv3 model, save it, and then load it to produce a prediction for a new shot [16].

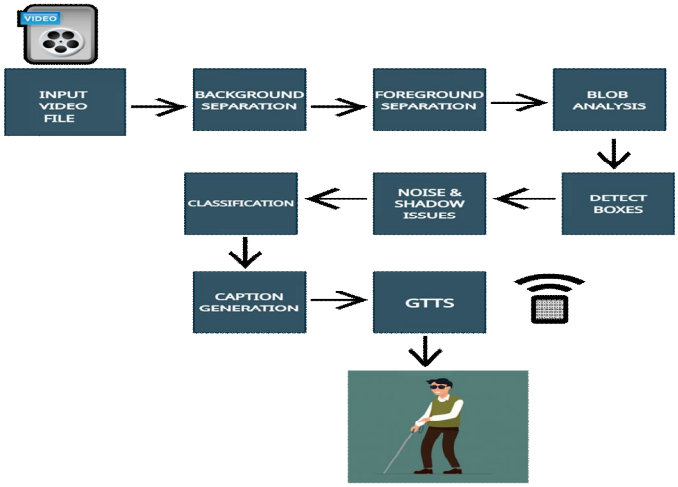


Fig. 1 Block Diagram of Object Detection with Captioning

Fig. 1 shows the block diagram of object detection with captioning where firstly, a video file or a camera-based live detection is given as the input. The background separation/subtraction process comes next. Backdrop subtraction is a technique for distinguishing the foreground from the background. A foreground filtration process is also carried out. Where the foreground is separated from the backdrop in preparation for subsequent processing. Blob analysis is the next step. Blob stands for Binary Large Object and refers to the binary image's connected pixels (represented in black and white colour). The adjective "large" refers to an entity of a definite size, whereas "little" binary objects are typically noise. The bounding boxes for each object are then determined. Noise and shadow difficulties are represented by the additional block. This block removes excessively small items that are considered noise, as well as shadows (if enabled). The bounding boxes that have been discovered are then classified using a class map.

B. Virtual Text Reader

This technology allows those who are blind or visually handicapped to read newspapers, books, periodicals, journals, and other publications. Fig. 2 shows the block diagram of virtual text reader.

1) Image acquisition: The input text picture is read and additional processing begins in this step. The visuals of the text are captured by the built-in camera. The image quality captured is determined by the camera used. Employing a web camera with a resolution of 2592x1944 and a resolution of 5MP or greater.

2) Image pre-processing: Colour to monochrome conversion, noise removal, edge recognition, warping and thresholding, and cropping are all included in this stage. Because many OpenCV methods require a grayscale image as an initial parameter, the image is transformed to grayscale. A dual filter is used to remove noise. For improved contour detection, Canny edge detection is applied to the grayscale image. The image's warping and cropping are done using the contours as a guide. This allows us to recognize and extract only the text-containing regions while removing the undesired backdrop. Finally, Thresholding is applied to the image so that it resembles a scanned document. This is done to allow the OCR to transform the image to text as quickly as possible.

©IJRASET: All Rights are Reserved

3) *Image to text conversion:* The flow of Text-To-Speech is depicted above. The image pre-processing modules and OCR are the initial block. It converts the pre-processed picture (.png) to a text file (.txt). Tesseract OCR is what we're utilizing.
4) *Text to speech conversion:* The voice processing module is the second block. The.txt file is converted to an audio output. A speech synthesizer named Festival TTS is used to convert the text to speech.
5) *Audio output:* Finally, speakers are used to generating output voice.

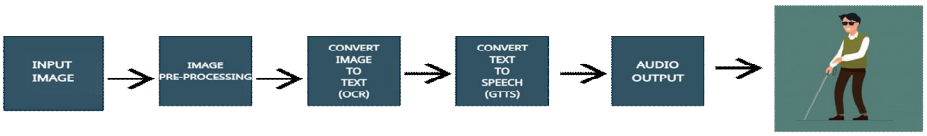


Fig. 2 Block Diagram of Virtual Text Reader

C. Live text Image to Voice

The previously outlined approach can be used to produce live text images to speech. Rather than using a previously saved image, we use a live image taken by the camera at the time. This method can be used to detect medication for visually impaired people. It can also be used to read brief passages of text.

IV. RESULTS

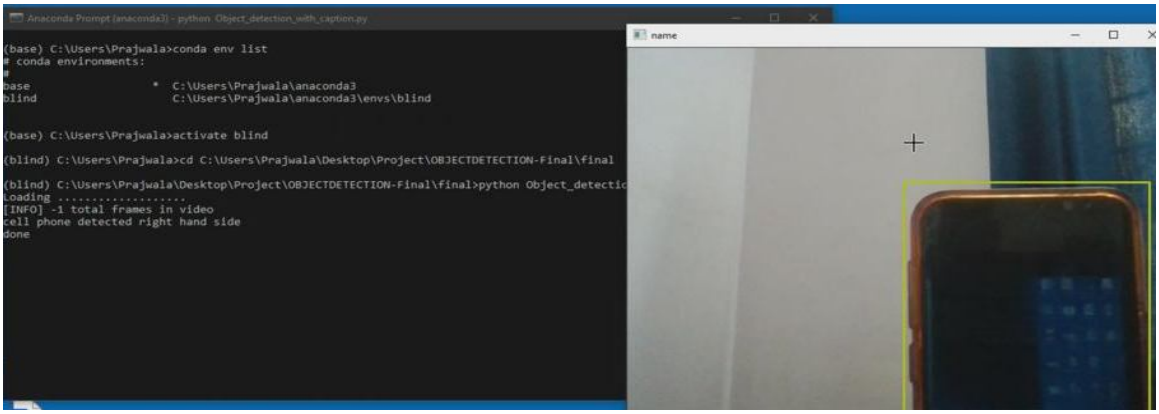


Fig. 3 Object Detection with Captioning

The outcome for one of the objects, as shown in Fig. 3, where the object detected is a cell phone and represented by a bounding box. “Cell phone detected right hand side” is the caption that is generated. When the object is placed on the left side, a caption such as "cell phone detected left hand side" is generated. In addition, an average of 90 or more objects can be identified.

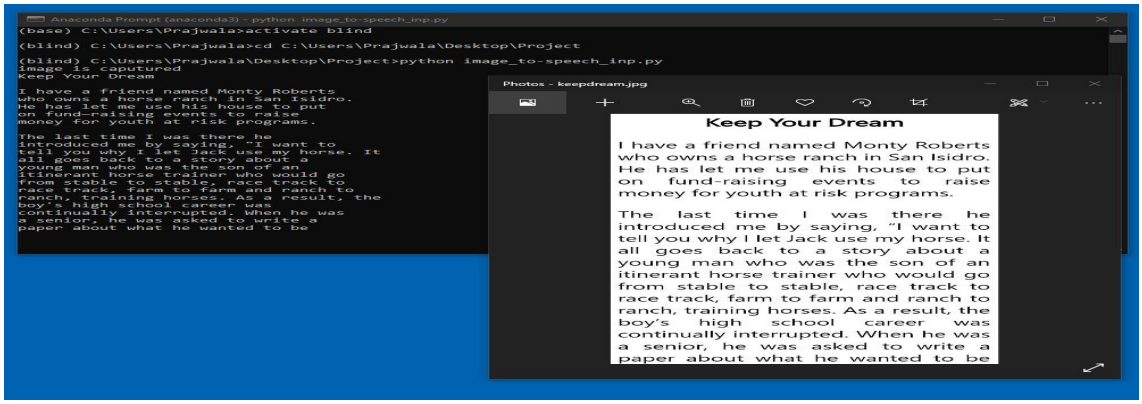


Fig. 4 Virtual Text Reader

The example of a virtual text reader is illustrated in Fig. 4 above. As an input file, a text image with the phrase "Keep Your Dream" is provided, and the message in the text image is displayed on the screen and read aloud. This can be used to read long paragraphs, novels, newspapers, journals, and other materials aloud.

©IJRASET: All Rights are Reserved

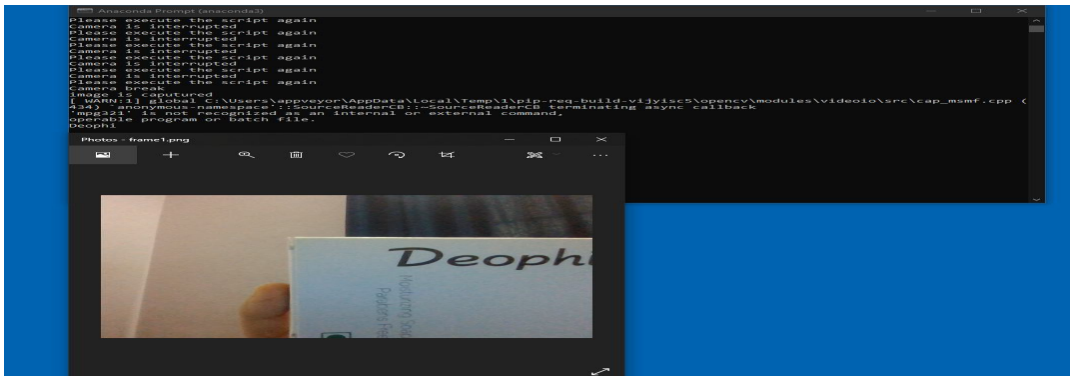


Fig. 5 Live Text Image to Voice

The camera captured a live text image, as illustrated in Fig. 5, and the corresponding word "Deophi" was displayed on the screen and read aloud. Text on pharmaceuticals, identify books, and other small text images can all be detected using the same method.

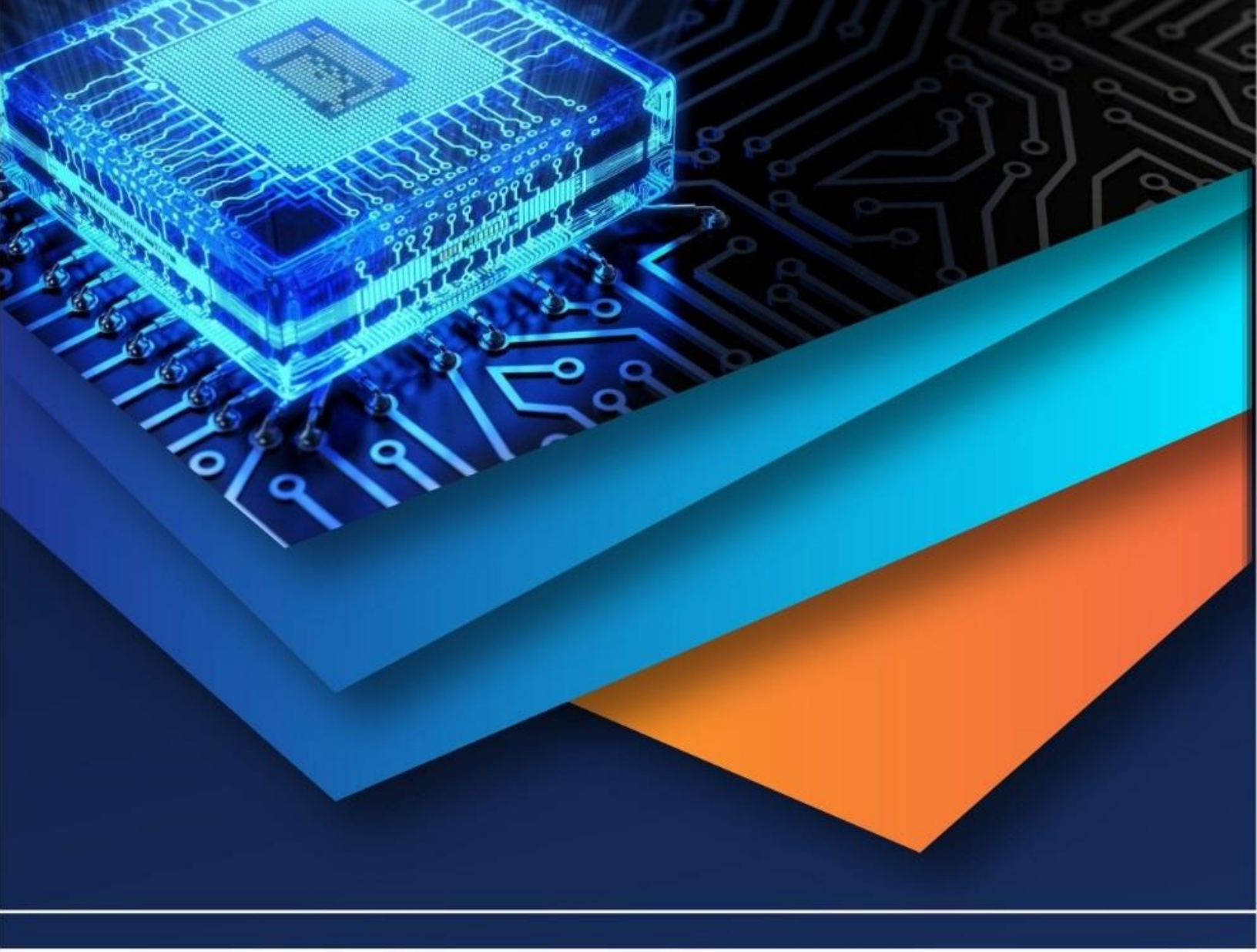
V. CONCLUSION

According to WHO, the number of visually impaired people of all ages worldwide is estimated to reach 285 million, with 39 million of them blind [17]. People aged 50 and up account for 82% of all blind people. Uncorrected refractive errors (43%) and cataract (33%) are the leading causes of visual impairment; cataract is the leading cause of blindness (51%) [18]. Our proposed plan enables this percent of the multitude to lead an independent and almost normal lifestyle and to ultimately aid them in situations where there's nobody around. The suggested concept uses live photos to detect objects and provide captions that can be turned to voice to help the visually impaired understand what is going on around them. Subsequently, this technique can also be used on text images, allowing a visually impaired individual to recognize text on drugs and other things with text. Furthermore, with the help of a virtual text reader, they would be able to obtain knowledge about current events as well as enjoyment.

REFERENCES

[1] W. Wei, "Small Object Detection Based on Deep Learning," 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), 2020, pp. 938-943.
[2] Ansar Hani, Najiba Tagougui, Monji Kherallah. Image Caption Generation Using A Deep Architecture, IEEE International Arab Conference on Information Technology (ACIT), 2019.
[3] Varsha Kesavan, Vaidehi Muley, Megha Kolhekar. Deep Learning Based Automatic Image Caption Generation, IEEE Global Conference for Advancement in Technology (GCAT) Bangalore, India, October 2019.
[4] Pabasara Jayawardhana, Amila Rathnayake. An Intelligent Approach of Text to Speech Synthesizers for English and Sinhala Languages, IEEE 2nd International Conference on Information and Computer Technologies, 2019.
[5] S. U. Nisa and M. Imran, "A Critical Review of Object Detection using Convolution Neural Network," 2019 2nd International Conference on Communication, Computing and Digital Systems (C-CODE), 2019, pp. 154-159.
[6] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, D. Lin, "Libra R-CNN: Towards Balanced Learning for Object Detection," Computer Vision and Pattern Recognition, pp:821-830,2019.
[7] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, K. Cho, "Augmentation for small object detection," Computer Vision and Pattern Recognition (CVPR), pp.1-15, Jun 2019.
[8] V. R. Prakash, Saran. An Enhanced Coding Algorithm for Efficient Video Coding. Journal of the Institute of Electronics and Computer, 1, 28-38, 2019.
[9] A. Lumini, L. Nanni, A. Codogno and F. Berno. Learning morphological operators for skin detection. Journal of Artificial Intelligence and Systems, 1, 60–76, 2019.
[10] B. Singh, L. Davis, "An Analysis of Scale Invariance in Object Detection – SNIP," Computer Vision and Pattern Recognition, pp.3578-3587, Oct 2018.
[11] Z. Cai, N. Vasconcelos, "Cascade R-CNN: Delving into High Quality Object Detection," Computer Vision and Pattern Recognition, pp.6154-6162, Oct 2018.
[12] N Komal Kumar, D Vigneswari, A Mohan, K Laxman, J Yuvaraj. Detection and Recognition of Objects in Image Caption Generator, IEEE 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019.
[13] Jayakumari J, Femina Jalin A. An Improved Text to Speech Technique for Tamil Language using Hidden Markov Model, IEEE 7th International Conference on Smart Computing & Communications (ICSCC), 2019.
[14] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5820628/
[15] https://blog.roboflow.com
[16] https://machinelearningmastery.com
[17] Emirates National Schools Student Paper
[18] He Cao, Lu Zhang, Liping Li, SingKai Lo. "Risk Factors for Acute Endophthalmitis following Cataract Surgery: A Systematic Review and Meta-Analysis", PLoS ONE, 2013

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)