SANJAY: Assistive Vision for the Blind

Prashant Kanade, Omkar Mangalpalli, Abha Ranade, Divya Raisinghani, Mayur Pawar

Vivekanand Education Society's Institute of Technology

IJASR 2021
VOLUME 4
ISSUE 3 MAY – JUNE **ISSN: 2581-7876**

Abstract: SANJAY is an assistant for the visually impaired. It is a "Captionbot for assistive vision" that provides a reading facility as well as an object detection tool for the blind. Disability is simply a mismatch between a person and his surroundings. People are invincible when they have the right resources. By effectively educating blind or visually disabled people of their environment, this project aids in addressing the challenges associated with visual disability. The project not only aids in the understanding of the environment but also gives the consumer a sense of freedom. The objects are detected and their names are displayed on a computer before being translated to expression. Our project uses YOLO (You Only Look Once) detection models for object detection after a thorough examination of previous detection algorithms such as Convolution Neural Network (CNN), Region dependent CNN (RCNN), Fast RCNN, Faster RCNN, and YOLO.

Keywords: visually impaired, captionbot, deep learning, assistive technology, object recognition

Introduction

Millions of people live in this world with incapacities of understanding the environment due to visual impairment. Although they can develop alternative approaches to deal with daily routines, they also suffer from certain navigation difficulties as well as social awkwardness. Computer vision, in particular, mobile vision is gaining popularity as an assistive technology for people with visual impairments. Algorithms and systems advancements are paving the way for modern and exciting applications. However, there is still a lack of knowledge about how a blind person would use a camera-based device. This knowledge is needed not only for designing good user interfaces, but also for correctly dimensioning, designing, and benchmarking a mobile vision device for these applications.

On the basis of computational and storage costs, this decision was taken after considering the drawbacks of all of the above detection techniques. The position and scale of the bounding boxes from the detection algorithm are used to estimate the 3D location of the items. This detection is active on the user's handheld computer or cell phone, which uses its camera and microphone for real-time input feed and audio. The sound is delivered to the user via speakerphone or earphones, depending on the user's preference. It is played at short intervals or when the recognized object differs from the previous one, whichever comes first.

Computer vision technologies, especially the deep convolutional neural network, have been rapidly developed in recent years. It is promising to use state-of-art computer vision techniques to help people with vision loss. Computer vision is an interdisciplinary research area that studies how machines can be programmed to interpret visual images or videos at a high level. It aims to automate tasks that the human visual system can perform from an engineering standpoint. Methods for collecting, extracting, analyzing, and interpreting digital images, as well as the extraction of high-dimensional data from the real world in order to generate numerical or symbolic knowledge, such as in the form of

Decisions are all examples of computer vision tasks. In this sense, understanding refers to the conversion of visual representations into interpretations of the environment that can be used to communicate with other thought processes and evoke effective action. Computer vision is a scientific discipline that studies the principle behind artificial systems that extract knowledge from images. Video loops, multiple camera views, or multidimensional data from a medical scanner are all examples of image data.

Over the years various platforms have been developed, one of such is Microsoft CaptionBot. It often responds with "I am not really confident" to the images. It also doesn't read the captions out loud. Experimental results show that this task still has better performance systems and improvement.

In the proposed paper a captionbot for assistive vision using a deep learning model has been put forth, which can take video or image inputs and these images can then be used to automatically generate captions in properly formed English sentences that can be read out loud to the visually impaired so that they can get a better sense of what is happening around them. In this project, we want to explore the possibility of using the hearing sense to understand visual objects. The sense of sight and hearing sense share a striking similarity: both visual object and audio sound can be spatially localized. It is not often realized by many people that we are capable of identifying the spatial location of a sound source just by hearing it with two ears. In our project, we build a real-time object detection and position estimation pipeline, with the goal of informing the user about the surrounding object and their spatial position using binaural sound.

Literature survey

Author Roberto Manduchi's paper [1] presents an experimental study of a sign-based wayfinding system that employs a mobile camera to detect unique color markers to allow exploration of an indoor environment without sight. In three practical indoor settings, eight blind volunteers of various ages were asked to use the device. According to their research, the narrow field of view of traditional mobile cameras is an obstacle to the system's practicality. Several participants found it difficult to use a handheld device and suggested that a camera mounted to their body would be preferred.

Authors Jinqiang Bai, Dijun Liu, Guobin Su, and Zhongliang Fu [2] propose a cloud and vision-based navigation system for blind people that consist of a helmet with attached cameras, an android-based smartphone and a web application. The system is based on a cloud computing platform that combines deep learning algorithms for object detection and recognition, OCR (Optical Character Recognition), speech processing, vision-based SLAM (Simultaneous Localization and Mapping), and route planning. Blind people use voice to communicate with the machine. Wi-Fi or 4G mobile communication technology is needed to connect the smartphone to the cloud platform. Vision test and navigation test were carried out to test the performance of the system. The results show that the proposed system can recognize 1000 different types of common objects with an accuracy of 99.9% for the RMB recognition dataset.

Author Quoc Khanh [3] proposes a virtual blind cane device for indoor use, which includes a mobile camera, a laser, and an inertial measurement unit (IMU). The type of obstacle is determined by examining the laser intersection histogram, and the distance to the obstacle is calculated using a combination of swing movement analysis and 3D laser point coordinates. A blind person may use the proposed device to determine the type of obstacle and its distance. Experimental results show that the obstacles can be classified accurately, and the distance to the

Obstacle has a small error. However, the system's drawback is that it is affected by bright lighting.

In the paper titled 'Real Time Text Detection and Recognition on Hand Held Objects to Assist Blind People' [4], authors S. Deshpande and R. Shriram propose a novel text detection and recognition algorithm that employs MSER feature (Maximally Stable External Regions) to extract text information from images and OCR (Optical Character Recognition) to identify localized text patterns. The paper describes a camera-based system that recognizes text on handheld objects and reads it out loud in audio format to assist visually disabled people. The results of experiments with various text patterns show that the proposed algorithm provides high accuracy for text detection.

Authors Ádám Csapó, György Wersényi, Hunor Nagy, and Tony Stockman [6] summarise recent advances in audio and tactile feedback-based assistive technologies aimed at the blind community along with the summary of the general capabilities offered by cutting-edge mobile platforms that can be used to support assistive solutions for the visually impaired. There are several areas and issues for research such as mode selection, where there isn't anything to advise designers about how to present information in the best possible way; type selection, little is also known about how data will be mapped to the chosen mode, such as via (direct) representation-sharing or (analogy-based) representation-bridging.

Methodology

OPENCV:

OpenCV is a computer science branch that focuses on how machines can be rendered to achieve high-level comprehension from visual images or videos.

Due to recent advancements in the field of Deep learning and Artificial intelligence, OpenCV has been able to surpass human vision in tasks related to object detection and labeling.

Thanks to largely available data of images we generate today that is then used to train and make computer vision better. Using Deep learning, we try to automate tasks that the human visual system can do i.e. the transformation of visual data into descriptions or text. This can be seen as combined results of statistics, image processing, deep learning model, and learning theory.

The main goal of computer vision is to determine whether image data contains some features or not. Scene reconstruction, video tracking, motion estimation, image restoration, and object recognition are some sub-domains of computer vision. In computer vision, the object is stored on basis of its attributes and properties further refinement and adjustments can be performed in order to improve the quality and speed of object identification. The input image goes through the following processes - image sampling and compression, segmentation, feature selection and extraction, image recognition and interpretation which is very much similar to human interpretation. Image recognition is used as object classification where the model is pre-train using several predefined classes of images. It can be used to identify signatures, fingerprint, digits, road signs etc. Images can be scanned to look for specific conditions such as to detect cancer cells in the medical field. Correct and fast interpretation is possible because of smaller areas of interest.

Convolutional neural networks are the best option for image classification because pattern matching is performed by moving filters across images. Overfitting is avoided because of Dropouts. Various CNN models have introduced ImageNet Large Scale Visual Recognition Challenge e.g. AlexNet, ZFNet, VGG-16, RESNET-50, MobileNet etc.

Neural Networks:

Neural Network is a series of neurons that recognize underlying relationships like a human brain. They are used in a variety of operations such as fraud detection, image processing, feature extraction, finance, etc. A "neuron" in a neural network is nothing but a mathematical function that collects and classifies information according to a specific architecture. Single neuron acts as a binary classifier. It takes a weighted sum of input and applies a non-linear activation function to give output. Perceptron is similar to multiple linear regressions. Multi-layer perceptrons are dense layers of interconnected neurons. Usually, binary cross-entropy is used to calculate loss between expected output and predicted output. Then a backpropagation algorithm is used to train the MLP model. MLP is used in the last stage of classification in CNN based models to classify input features into objects. CNN is used as a feature extraction whereas MLP is used as a classifier.

Deep Learning:

Deep Learning is Hierarchical Feature Learning. This hierarchy of concepts allows the computer to exploit the patterns and concepts by building them out of simpler ones. If we draw a graph to show how these concepts are built the graph becomes deep and has many layers. The higher- level layers are built by combining lower-level layers. In Image processing layers may identify edges, while higher layers may identify the concepts relevant to a human such as digits or letters or faces Various architectures such as deep neural networks, convolutional neural networks, recurrent neural networks are used in various fields like computer vision, machine vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection, board game programs and Natural Language Processing - Text Generation, Text Analysis, Text Translation, Chatbots. Convolutional neural networks play an important role in feature extractions from an image. Then these features are used to generate text using a deep learning model.

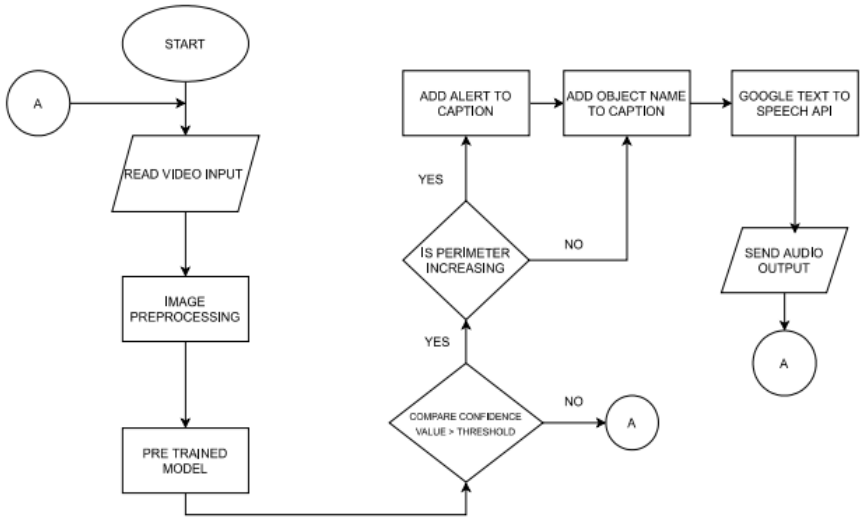
YOLO (You Only Look Once):

YOLO is an object detection algorithm. The object detection task entails locating and classifying specific objects on an image. Previous approaches, such as R-CNN and its variants, used a pipeline to execute this role in several stages. Since each individual component must be trained separately, this can be slow to run and difficult to optimize. YOLO achieves these things with only one neural network. The input frame is divided into a grid of cells. One grid cell is said to be "responsible" for predicting each object that appears in the picture. Ultimately, we aim to predict a class of an object and the bounding box specifying object location. Each bounding box can be described using four descriptors:

1. Center of the box (bx, by)
2. Width (bw)
3. Height (bh)
4. Value c corresponding to the class of an object

Conceptual architecture

Our app takes input from a mobile camera. This input is fed to the YOLO model then the model generates audio output based on objects in the video. If the perimeter of the object is increasing then the user gets an alert about proximity through audio. The perimeter for each object is calculated by comparing the current frame with the previous frame. The output of the model is a confidence score for each object. If any object falls below the threshold value is ignored, only objects having a confidence score above threshold are sent for audio feedback.



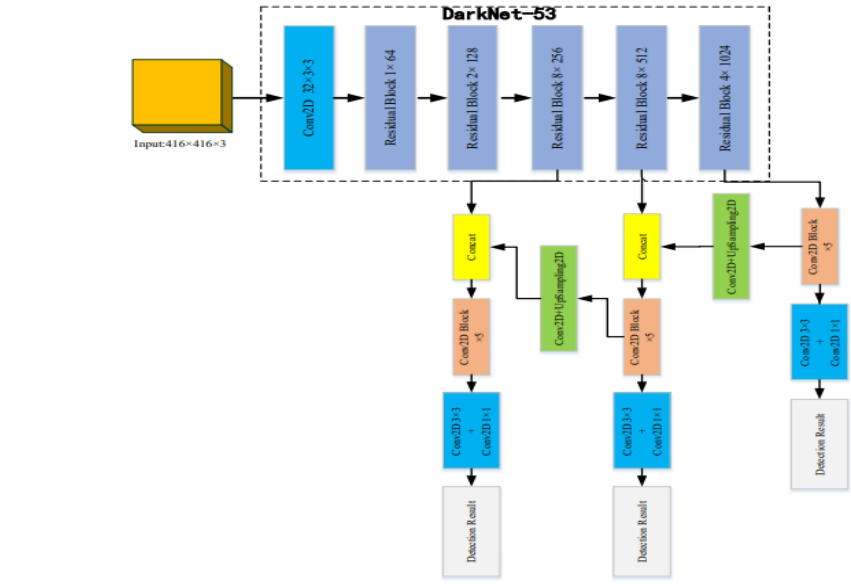
CONCEPTUAL ARCHITECTURE OF THE SYSTEM

Analysis of frame with YOLO (You Only Look Once):

Following are the steps involved in Network Prediction process of YOLO algorithm:

1. The frames with a size of 416 * 416 are first uploaded to the Darknet-53 network. A feature map of size 13 * 13 is obtained after several convolutions, and then 7 times by 1 1 and 3 * 3 convolution kernels are processed to achieve the first class and regression bounding box prediction.
2. The scale 13 * 13 feature map is processed 5 times with 1 * 1 and 3 * 3 convolution kernels, followed by 2 times the upsampling layer and stitching to the size on the 26 * 26 feature map. To obtain the second group and regression bounding box estimation, the new function map of size 26 * 26 is processed seven times using 1 * 1 and 3 * 3 convolution kernels.

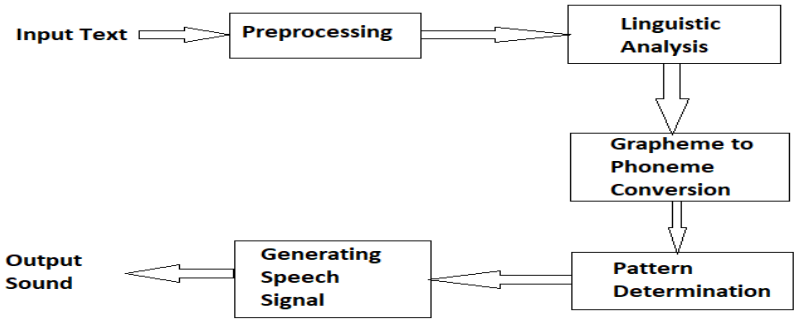
3. The scale of a new function map is 26 x 26. To begin, we process 5 times with 1 * 1 and 3 * 3 convolution kernels, then perform a double upsampling operation and stitch it onto a 52 * 52 feature map. To obtain the third group and regression bounding box estimation, the function map is processed seven times using 1 * 1 and 3 * 3 convolution kernels.



Input Frame | Center Detection | Marking rectangle with Width and height | Printing object name from list

Conversion of output text to voice using gtts (Google - Text to Speech):

To output the text generated by our model as a voice through device speaker, we used Google's text-to-speech conversion library. This is one of the best engines to give audio feedback by converting inputted text data to sound. The working of this module can be easily understood by the following diagram:

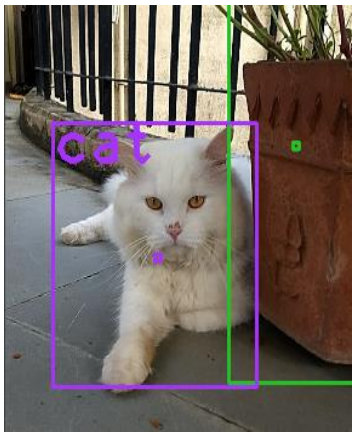


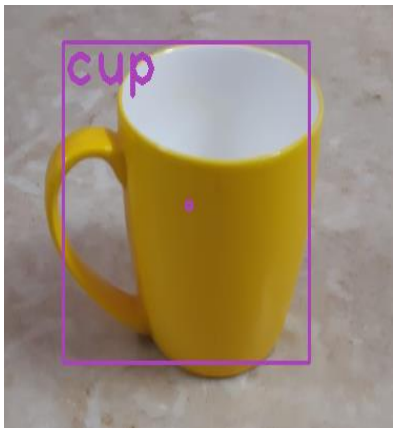
Implementation setup

Users can interact with the app by simply pointing a mobile camera towards the object. The app takes live video as an input and detects various objects and generates a caption to describe an image and then this caption is converted to audio by google text to speech API. App also detects if an object is moving closer to the frame and gives audio feedback to the user. Users get real time audio feedback about his/her surroundings.

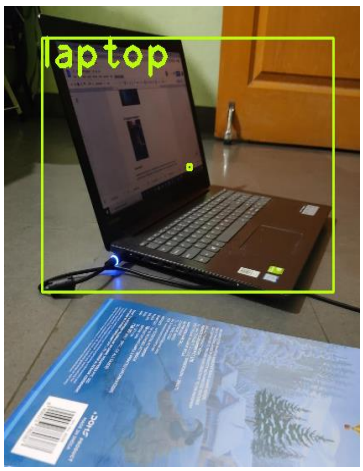
Results and Analysis

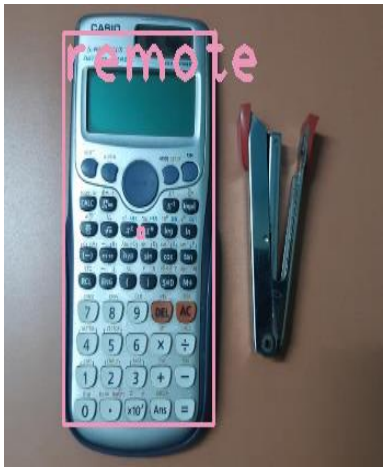
1) Accuracy of Detection :





2) Multiple Object Detection:





3) Incomplete Detection:

4) Negative feedback:

The project achieved more rigorous detection under the use of YOLO than any other detection algorithm tested, resulting in satisfactory results and assisting the project in being an aid for affected users. On the mobile device, the model ran smoothly and successfully, offering an inexpensive and effective forum for harnessing and instilling the benefits of artificial intelligence in people's everyday lives. This model gave pretty accurate and fast results over input frames and was perfect to use for our project since the exact location of objects was getting pinpointed. The model performed well on objects which were partially present in the frame. It was observed that the model sometimes tends to focus only on one object and ignores the other objects if multiple objects are present in a frame and if relative sizes of objects vary to a very large extent.

Conclusion

Captionbot for assistive vision is a highly effective and robust device designed to aid visually impaired people to identify objects in their surroundings and know their proximity. The basic outline of this system consists of taking real-time video input from a camera, object recognition using deep neural networks, and text to speech conversion for audio output through speaker or headphones. Following a thorough review of various algorithms such as Convolution Neural Network (CNN), Region based CNN (RCNN), Fast RCNN, Faster RCNN, and YOLO (You Only Look Once), our system has employed the YOLO network for object detection. Since the other algorithms employ a pipelined approach, they are time consuming and difficult to optimize. On the other hand, YOLO accomplishes the same task faster by using a single neural network. The mobile application is easy to use and can be accessed by anyone owning an android smartphone. Assistive technology for the visually impaired is not a luxury, it is a necessity. This project employs the concepts of computer vision and deep learning to build an inexpensive and handy app that speaks what it sees in order to help blind people gain a better understanding of their surroundings.

Future scope

The system that we have developed is a mobile application. An IOT device that can be attached to the body or clothes of a blind person may be a more convenient and viable solution as it will enable the person to keep his or her hands unoccupied. For instance, using the same underlying technology, smart glasses can be developed by mounting a camera on the glasses of a visually impaired person and giving audio description through the headphones. With the required enhancements the system can be used to detect text written on objects in multiple languages and give appropriate audio output in the recognized language. The system can also be enhanced to give output in form of complete sentences with minimal grammatical errors, instead of a few words.

References

1. Manduchi R. (2012) Mobile Vision as Assistive Technology for the Blind: An Experimental Study. In: Miesenberger K., Karshmer A., Penaz P., Zagler W. (eds) Computers Helping People with Special Needs. ICCHP 2012. Lecture Notes in Computer Science, vol 7383. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-31534-3_2

2. Bai, Jinqiang & Liu, Dijun & Su, Guobin & Fu, Zhongliang. (2017). A Cloud and Vision-based Navigation System Used for Blind People. 1-6. 10.1145/3080845.3080867.
3. Dang QK, Chee Y, Pham DD, Suh YS. A Virtual Blind Cane Using a Line Laser-Based Vision System and an Inertial Measurement Unit. Sensors (Basel). 2016;16(1):95. Published 2016 Jan 13. doi:10.3390/s16010095
4. S. Deshpande and R. Shriram, "Real time text detection and recognition on hand held objects to assist blind people," 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, India, 2016, pp. 1020-1024, doi: 10.1109/ICACDOT.2016.7877741.
5. Ashraf, M., Hasan, N., Lewis, L., Hasan, M., & Ray, P. (2016). A Systematic Literature Review of the Application of Information Communication Technology for Visually Impaired People. International Journal of Disability Management, 11, E6. doi:10.1017/idm.2016.6
6. Csapó, Á., Wersényi, G., Nagy, H. et al. A survey of assistive technologies and applications for blind users on mobile platforms: a review and foundation for research. J Multimodal User Interfaces 9, 275–286 (2015). https://doi.org/10.1007/s12193-015-0182-7
7. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788.doi: 10.1109/CVPR.2016.91

A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

8. Farhoodfar, Avid. (2019). Machine Learning for Mobile Developers: Tensorflow Lite Framework.
9. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. The MIT Press.
10. Umberto Michelucci. 2019. Advanced Applied Deep Learning: Convolutional Neural Networks and Object Detection (1st. ed.). Apress, USA.