International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (01), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Real Time Object Detection for Blind People

N.Saranya[1], M.Nandinipriya[2], U.Priya[3]
[1,2,3] *Assistant Professor, Department of Electronics and Communication Engineering,*
Bannari Amman Institute of Technology, Sathyamangalam, Erode.(India)

ABSTRACT
Good vision is a precious gift but unfortunately loss of vision is becoming common now a days. To help the blind people the visual world has to be transformed into the audio world with the potential to inform them about objects as well as their spatial locations. Objects detected from the scene are represented by their names and converted to speech. The blind people's spatial locations are encoded into 2-channel audio with the help of 3D binaural sound simulation. Video is captured with a portable camera device on the client side, and is streamed to the server for real-time image recognition with existing object detection models. The 3D location of the objects is estimated from the location and the size of the boundary boxes from the detection algorithm. Then, a 3D sound generation application based on unity game engine renders the binaural sound with locations encoded.

Keywords: Computer vision technique, Droidcam, Morphological algorithm, Object tracking, YOLO

I. 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 suffer from certain navigation difficulties as well as social awkwardness. For example, it is very difficult for them to find a particular room in an unfamiliar environment. And blind and visually impaired people find it difficult to know whether a person is talking to them or someone else during a conversation.

Computer vision technologies, especially the deep convolutional neural network, have been rapidly developed in recent years. It is promising to use the state-of-art computer vision techniques to help people with vision loss. This paper aims at exploring 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 at identifying the spatial location of a sound source just by hearing it with two ears. The aim of the work is to guide the blind people through the output of processor or controller by voice to navigate them. The methodology of this work includes Object Extraction, Feature Extraction, Object Comparison

There exists multiple tools to use computer vision technologies to assist blind people. The mobile app "TapTapSee" uses computer vision and crowd sourcing to describe a picture captured by blind users in about 10 seconds. The Blind sight offers a mobile app Text Detective featuring optical character recognition (OCR) technology to detect and read text from pictures captured from the camera. However, these products were not

focusing on enabling general visual sense for blind people and did not use the spatial sound techniques to further enhance the user experience.

II. EXISTING METHOD

2.1 YOLO MODEL

YOLO can correctly detect objects, such as chair, within a range about 2-5 m away but the objects that are outside this range are either unrecognized or misclassified. The second issue reported by the blind user is the blocking of ambient sound by using earbuds. The third issue reported by

the blind user is "information overload" when the system is trying to notify user of multiple objects at the same time. This can be solved by delayed notifications.

2.2 POINT DETECTION AND TRACKING

In this method, the extract the prominent feature points from each target object is obtained and then uses a particle filter based approach to track the feature points in image sequences based on various attributes such as location, velocity and other descriptors. They used rectangular bounding box for object representation. But this algorithm may not successfully track feature points with different velocities. Hence this algorithm needs more flexible object representation and also they used static camera for capturing the video

2.3 SEGMENTATION BASED DETECTION AND TRACKING

A unified framework for both single and cross camera tracking with affinity constraints using graph matching was proposed. In this method, they mainly dealt with the problem of existence occlusion in single camera scenario & the occurrence of transition in cross camera scenario and also they consider the data association method in handling occlusion. They consider the tracklet association problem as graph matching with affinity constraints and leverage both person wise and part wise attribute for similarity measurement between tracklets to overcome the uncertainty and noise. The crucial problem caused by cross camera tracking lies in the drastically increasing the data.

Instead of using region proposal method, YOLO model divides an image into grid. Each grid cell predicts B bounding boxes, and boxes confidence scores for the prediction and detect if a class falls in the boxes.

After detecting the type of objects in a video frame, the next step is to obtain the depth or distance of the detected object from the user. Make two separate attempts to this problem. In initial attempt, we use a Microsoft Kinect as the video camera device in our pipeline. Kinect has the benefit of capturing real-time depth map together with the RGB image. After detecting the object in the RGB image and its corresponding bounding box, simply use the average depth of the bounding box area as the distance. There exist other more portable depth cameras, for example the Zed camera with relies on depth estimation from stereo vision. However, such camera generally requires high computation resource (GPU) to estimate depth from stereo image.

A pipeline is developed that enables us to communicate quickly. The server decodes it and use trained object detection engine to return detected items. The server then sends that information back to the client, which

triggers the Unity-based stereo generator to play the 3D sound. During the implementation, we find that the communication between our personal computer and Rye machine takes a few milliseconds to transfer each video frame. Also the performance of the Rye machine is not stable due to the mass occupancy of GPU.

2.4 PROBLEM IN EXISTING METHOD

YOLO outputs the top classes and their probability for each frame. We take any probability above 20% as a confident detection result. The algorithm also has to decide whether to speak out a detected object and at what time. Obviously it's undesirable to keep speaking out the same object to the user even if the detection result is correct. It's also undesirable if two object names are spoken overlapping or very closely that the user won't be able to distinguish.

To solve the first problem. For example, if a person is detected in the first frame and is speaking out, the program will not speak out "person" again until after five seconds. This is only a sub-optimal solution since it does not deal with multiple objects of the same class. Ideally, if there are two persons in the frame, the user should be informed about the two person, but he does not need to be informed about the

same person continuously. One possible improvement, which we are still working on, is to track the object using overlapping bounding box between frames. The second issue reported by the blind user is the blocking of ambient sound by using ear buds. However, this can be solved by using bone conduction earphones, which leave ears open for hearing surrounding sounds. The third issue reported by the blind user is "information overload" when the system is trying to notify user of multiple objects at the same time. This can be solved by delayed notifications.

III. PROPOSED METHODOLOGY

This paper gives a criterion for designing a time-efficient cascade that explicitly takes into account the time complexity of tests (as evaluated by computer run time) including the time for pre-processing. We design a greedy algorithm to minimize this criterion. Finally, we illustrate our method on the task of image detection in snap. This gives a detection algorithm that runs at 0.025 seconds per 320×240 image. This is a speed up factor of 2.5 compared to our previous image detector. It gives a real time system which can be used for applications to help the blind and visually impaired. Among all 20 classes of the existing model, we choose the following classes to inform the user: "chair", "table", "laptop", "TV monitor".

3.1VIDEO TO FRAME CONVERSION

Frames can be obtained from a video and converted into images using MATLAB function. The original format of the video that has be used as an example is .gif file format and it is converted into an avi format video.

3.2 PRE-PROCESSING

In imaging science, image processing is processing of images using mathematical operations , the output of which may be either an image or a set of characteristics or parameters related to the image. Closely related to image processing are computer graphics and computer vision.

3.3 BACKGROUND FRAME INITIALIZATION

The main purpose of foreground/background segmentation, a basic process of a computer vision application system, is to extract some interesting objects (the foreground) from the rest (the background) of each video frame in a video sequence background subtraction is a popular foreground background segmentation approach, which detects the foreground by thresholding the difference between the current video frame and the modelled background in a pixel-by-pixel manner. The correctness of the modelled background is usually affected by three factors: illumination changes, dynamic backgrounds, shadows

3.4 BACKGROUND SUBTRACTION

Identifying moving objects from a video sequence is a fundamental and critical task in many computer-vision applications. A common approach is to perform background subtraction, which identifies moving objects from the portion of a video frame that differs significantly from a background model. There are many challenges in developing a good background subtraction algorithm. First, it must be robust against changes in illumination. Second, it should avoid detecting non-stationary background objects such as moving leaves, rain, snow, and shadows cast by moving objects. Finally, its internal background model should react quickly to changes in background such as starting and stopping of vehicles.

3.5 SEGMENTATION

In computer vision, image segmentation is the process of partitioning a digital image into multiple segments. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image (see edge detection). Each of the pixels in a region are similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different with respect to the same characteristic when applied to a stack of images, typical in medical imaging, the resulting contours after image segmentation can be used to create 3D reconstructions with the help of interpolation algorithms like Marching cubes.

3.6 MORPHOLOGICAL FILTERING

Morphological image processing is a collection of non-linear operations related to the shape or morphology of features in an image. Morphological operations rely only on the relative ordering of pixel values, not on their numerical values, and therefore are especially suited to the processing of binary images. Morphological operations can also be applied to greyscale images such that their light transfer functions are unknown and therefore their absolute pixel values are of no or minor interest. Morphological techniques probe an image with a small shape or template called a structuring element. The structuring element is positioned at all possible locations in the image and it is compared with the corresponding neighborhood of pixels.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (01), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

3.7 OBJECT EXTRACTION

Object detection is the process of finding instances of real-world objects such as faces, bicycles, and buildings in images or videos. Object detection algorithms typically use extracted features and learning algorithms to recognize instances of an object category. It is commonly used in applications such as image retrieval, security, surveillance, and automated vehicle parking systems.

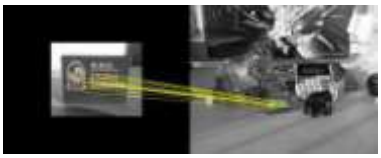

Fig1 Feature based object detection

Detecting a reference object (left) in a cluttered scene (right) using feature extraction and matching. RANSAC is used to estimate the location of the object in the test image. Local features and their descriptors are the building blocks of many computer vision algorithms. Their applications include image registration, object detection and classification, tracking, and motion estimation. These algorithms use local features to better handle scale changes, rotation, and occlusion. Computer Vision System Toolbox algorithms include the FAST, Harris, and Shi & Tomasi corner detectors, and the SURF and MSER blob detectors. The toolbox includes the SURF, FREAK, BRISK, LBP, and HOG descriptors.

From the stream, a single snap is took by using the function “snapshot” which is an inbuilt function. Major part of snapshot function is to take a snap of an image from the stream. The snapped image is directed into a variable by saving it which will be named. The path also given where the image have to save.

Video steam is paused for few seconds or minutes as mentioned. And that will be pause by using the inbuilt function “pause” which hold the video stream for mentioned timing. Within the duration the snapped image will be stored into the variable.

Fig 2 Object detection

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (01), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

3.8 IMAGE IDENTIFICATION

3.8.1 FIRST IMAGE

To successfully detect surrounding objects, we investigate several existing detection systems that could classify objects and evaluate it at various locations in an image. Using DroidCam application, an image is obtained by taking snapshot with some time interval. The image is stored into a folder through folder path by using the inbuilt function "imwrite". This will save the image to the specified variable as mentioned before the path.

Fig.3 First Image without object

3.8.2 SECOND IMAGE

After detecting the first image in a video frame, next step is to find the second image with object. Two separate attempts are handled in this problem.

In our initial attempt, we use a DroidCam wireless application as the video camera device in our pipeline. DroidCam has the benefit of capturing real-time RGB image. The second image also saved separately into a variable using unbuilt function. Same inbuilt function "imwrite" is used and also path is specified. For obtaining second image same pause function is used. But the time duration may vary. Both first and second images are stored in different variables but in same directory which help to perform further operations. While retrieving the second image preview of camera is not used. Only on obtaining the first image preview of camera is used.

Video steam is paused for few seconds or minutes as mentioned. And that will be pause by using the inbuilt function "pause" which hold the video stream for mentioned timing. Within the duration the snapped image will be stored into the variable.

Fig.4 Second Image with object

3.9 IDENTIFICATION OF OBJECT

As the images are already loaded for the video stream, they are retrieving from the directory by using the function "imread" which will save the image into another variable. Using the same inbuilt function two images were read. The size of the images is changed using imresize. X-factor and Y-factor get changes on applying imresize. Size will be specified.

First and second images are converted to grayscale which means the color image RGB is converted into grayscale. RGB images are 32-bit which is hard for manipulation so it is converted to grayscale. Where grayscale images are 16-bit, which easier for manipulation while compared to RGB images.

Binary images are much easier for manipulation. It is a black and white images which encloses zeroes and ones. Both grayscale images are subtracted which eliminates the same portions in both images. Only contrast portions are obtained which is called object extraction.

Extracted object is analyzed based on its size. The size varies for every object. Existing methodology uses various algorithms to extract the object but in this work subtraction plays major role. Extracted object is stored separately in a variable named. Line width, marker size, X-coordinates, Y-coordinates, intersection points are analysed. Through those the object will be decided.

This new time-efficient text detector cascade runs 2.5 times faster. The time cost of a classifier consists of the feature calculation time and the decision time of the classifier. Many applications employ some kind of preprocessing to save time for feature calculation.

3.10 OBJECT DETECTION USING BACKGROUND SUBTRACTION

To obtain background subtraction, the background has to model first. Then, the incoming frame is obtained, and subtract out from the background model. With the background model, a moving object can be detected. This algorithm is called as "Background Subtraction". The efficiency of a background subtraction technique correlates with three important steps: modelling, noise removal and data validation.

Background modeling, is the backbone of the Background Subtraction algorithm. Background model defines the type of model selected to represent the background, and the model representation can simply be a frame at time (t-1) formula such as the median model. Model Adaption is the procedure used for adjusting the background changes that may occur in a scene. Noise removal is a procedure that eliminates noise in the scene.

Data validation is involved with the collection of techniques to reduce the misclassification of pixels. A Gaussian mixture model (GMM) was proposed for the background subtraction in Friedman and Russell, and efficient update equations are given in Stauffer and Grimson. In Power and Schoonees, the GMM is extended with a hysteresis threshold. This method uses a Gaussian probability density function to evaluate the pixel intensity value. It finds the difference of the current pixels intensity value and cumulative average of the previous values. So it keeps a cumulative average (μ) of the recent pixel values. If the difference of the current images pixel value and the cumulative pixel value is greater than the product of a constant value and standard deviation then it is classified as foreground.

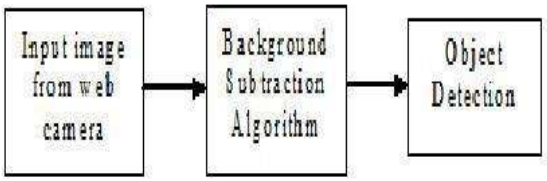

Fig.5.System block diagram

3.11 OBJECT TRACKING BASED ON COLOURED OBJECT

Object tracking means identifying & following same object in sequences of video frames. Camera is used as input sensors to acquire frames to form the video. The acquired video may have some noise. To remove noise from captured frames noise reduction technic is used to improve the image quality, to detect moving object, based on colour of the moving object in frame. Extraction of objects from frame using the different features is known as object detection. Every object has a specific feature based on its shape.

Applying background extraction algorithm, the object in each frame can be extracted out. The camera is capturing 30fps. The implementation is initially performed on matlab and various methods for object tracking are tested. The process of indicating the moving object in sequence of frames is known as tracking. This tracking can be performed by using the feature extraction of objects and detecting the objects in sequence of frames.

We are tracking the object are on basic colour RGB, to be detected object in frame we differentiate gray scale input image frame with coloured image frame to indicate coloured objet in video Rectangular Bounding Box. A rectangular coloured bounding box is plotted around the foreground objects detected from GMM based Background subtraction. By using the dimensions of rectangular bounding box, a centroid is plotted. The position of the centroid is stored & object is bounded in box.

FEATURE SET

- First order differential features calculated in blocks.
- Histogram features of intensity and gradient.
- Edge linking features.

3.12 DATA STREAMING

This work is based on a platform that is capable of processing real-time image. Thus, it is required to have a powerful GPU that could give feedback in no time. Considering the computational cost and performance, we initially use rye machine provided by Stanford as our prototype's server machine. A pipeline is developed that enables us to communicate quickly. A program in local machine extracts raw image from a camera (e.g. Kinect), encodes it into a string and sends through a client to a server running on the Stanford Rye machine. The server decodes it and use trained object detection engine to return detected items. Based on our initial platform, we switch to local platform that is more efficient. In this platform, the evnironment is captured by a portable camera

and transfers through HD video link directly to the computer through Droirdcam wireless application. The server detects objects, sends information directly to the unity sound generator and plays the binaural sound. In particular, the environment picture is captured by a portable GoPro Hero 3 at 30 frames 1080p resolution.

The video is live streamed through the HD video link to the computer server as shown in Figure

6. The HD video link could transfer a high resolution image within 2 miles in 1 millisecond. The object detection engine predicts objects in the stream. The algorithm could process a single image frame at a speed of 4-60 frames/second depending on the image size we send to the engine.

The outputs are sent to unity sound generator and the generated sounds are played through wireless earbuds. During the implementation, the platform is capable of processing all captured live stream at a minimum speed of 30 frames per second at 1080p resolution. 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 at identifying the spatial location of a sound source just by hearing it with two ears.

IV. RESULT AND DISCUSSION

To present the results to the user in a reasonable manner, the algorithm also has to decide whether to speak out a detected object and at what time. Obviously it's undesirable to keep speaking out the same object to the user even if the detection result is correct. It's also undesirable if two object names are spoken overlapping or very closely that the user won't be able to distinguish. The cool-down-time is assumed to be five seconds for each class. For example, if a person is detected in the first frame and is spoken out, the program will not speak out "person" again until after five seconds. This is only a sub-optimal solution since it does not deal with multiple objects of the same class. Ideally, if there are two persons in the frame, the user should be informed about the two person, but he does not need to be informed about the same person continuously. One possible improvement is to track the object using overlapping bounding box between frames. To solve the second problem, a delay of half a second between any spoken classes can be used.

The prototype we build successfully recognizes visual objects and presents the detection information as 3D sound, giving the user a sense of "augmented reality". However, the prototype suffers from the following limitations. First, it is common for user to focus on certain object from afar and navigate to a location close to the object. In this task, the user need a consistent instruction of the target object from approximately 10 m away to only 20 cm away. That impose a very high requirement to the object detection model. This work can correctly detect objects, such as chair, within a range about 2-5 m away. Objects that are outside this range are either unrecognized or misclassified.

One approach to solve this issue is to incorporate training images with greater scale ranges (e.g., include chair picture captured from 20 cm away and 10 m away). However, it may be difficult for object detection models to classify the object from a picture of extreme scale (too close or too far). Another approach to solve this is to use object tracking algorithm to track the object (e.g. a chair) once the user have identified as the target.

The second reported by the blind user is the blocking of ambient sound by using earbuds. However, this can be solved by using bone conduction earphones, which leave ears open for hearing surrounding sounds. The third

issue reported by the blind user is "information overload" when the system is trying to notify user of multiple objects at the same time. This can be solved by delayed notifications. For example, the system can sequentially notify the user of the object from left to right. However this solution requires the user stands still while playing the sounds.

Moreover, blind people usually do not want to know every objects in his "eyesight", but instead want to know objects that are pertinent to their immediate need. For example, they may want to find a particular room in a building, or find food and drinks during a conference.

In this regard, the system should have three modes: exploration mode where users are notified with every detected objects, search mode where the system only notify users of the object they are looking for, and navigation mode where only the target object and obstacle objects are notified to users in real time. In sum, extensive work is required to analyze users need if one would like to stem from this prototype to a really helpful assistive product.

V. CONCLUSION

In this prototype, we investigate the need from blind and visually impaired people. Base on the impetus of the CNN, we develop a blind visualization system that helps blind people better explore the surrounding environment. A portable and real time solution is provided in the work. We present a platform that utilizes portable cameras, fast HD video link and powerful server to generate 3D sounds.

By using morphological algorithm, the solution could perform accurate real time objective detection with 1080P resolution. A prototype for sensory substitution (vision to hearing) is established in the work. Through this work, we hope to demonstrate the possibility of using computer vision techniques as a type of assistive technology.

The training time for the time-efficient cascade was more than ten times longer than our previous method. But this computation is off-line and so is not significant. This will definitely navigate the blind people to detect the obstacles in front of them. Even there are plenty of techniques are available to guide them but this work efficiently help the blind people to guide and navigate them.

REFERENCES

[1.] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, pages 91–99, 2015.

[2.] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, pages 1440–1448, 2015.

[3.] B.Yu, L.Xu, and Y.Li, " Bluetooth Low Energy (BLE) Based Mobile Electrocardiogram Monitoring System," in Proc. of IEEE Int. Conf. on information and automation .pp,763-767,june,2014.

[4.] Jizhong Xiao, Kevin Ramdath, Manor Iosilevish, Dharmdeo Sigh, and Anastasis Tsakas. A low cost outdoor assistive navigation system for blind people. In Industrial Electronics and Applications (ICIEA), 2013 8th IEEE Conference on, pages 828–833. IEEE, 2013.

[5.] Tadas Naltrusaitis, Peter Robison, and Louis-Phileppe Morency, 3D Constrained Local Model for Rigid and Non-Rigid Facial Tracking (CVPR), 2012.

[6.] David Brown, Tom Macpherson, and Jamie Ward, Seeing with sound exploring different characteristics of a visual-to-auditory sensory substitution device. Perception, 40(9):1120–1135, 2011.

[7.] W.J.Li,Y. L.Luo, Y. S. Chang, and Y. H. Lin, "A wireless blood pressure monitoring system for personal health management," in Proc.32nd Annu. Int. Conf. IEEE EMBS, vol. 1, pp. 2196-2199, 2010.

[8.] Bonato, P., Wearable Sensors and Systems. IEEE Eng.Med. Biol.,volume 29(3) , page no 25–

[9.] May year 2010. Chen X, Ho CT, Lim ET, Kyaw TZ ,Cellular Phone based online ECG Processing for Ambulatory and Continuous Detection. Computers in Cardiology, volume 34, page no 653-656, year 2008.

[10.] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S.Wong and R.Young. "ICDAR 2003 Robust Reading Competitions", In 7th International Conference on Document Analysis and Recognition- 2003.

[11.] P. Viola and M. Jones, "Fast and Robust Classification using Asymmetric AdaBoost and a Detector Cascade", In Proc. of NIPS01, 2001

[12.] Aliverti, R Dellaca, P. P elosi, D.Chiumello, A. P edotti, and Gattinoni, "Optoelectronic plethysmography in intensive care patients," Am. J. Respir. Crit.