UMEED: Deep learning based walking and reading assistant for visually impaired people

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Technology

In

Computer Science and Engineering

Submitted by

Riya Goyal

Roll No. 801632042

Under the supervision of:

Dr. Parteek Bhatia

Associate Professor, CSED Department.

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY
PATIALA – 147004

June 2018

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*UMEED: Deep learning based walking and reading assistant for visually impaired people*", in partial fulfillment of the requirements for the award of degree of Master of Technology in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Parteek Bhatia* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Riya Goyal

(Riya Goyal)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

Parteek Bhatia

(Dr. Parteek Bhatia)

Associate Professor, CSED

ACKNOWLEDGEMENTS

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life.

This work would not have been possible without the encouragement and able guidance of my supervisor **Dr. Parteek Bhatia**. I thank my supervisor for their time, patience, discussions and valuable comments. Their enthusiasm and optimism made this experience both rewarding and enjoyable. Their discipline and sincerity towards work, teaches sincerity is more important than seriousness in life.

I am equally grateful to **Dr. Sanmeet Kaur**, Assistant Professor, Computer Science & Engineering Department, a nice person, an excellent teacher and a well – credited researcher, who always advised me with his valuable suggestions.

I will be failing in my duty if I don't express my gratitude to **Dr. S.S. Bhatia,** Senior Professor and Dean of Academic Affairs, Thapar Institute of Engineering and Technology, for making provisions of infrastructure such as library facilities, computer labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in the field.

I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love and affection, which made my stay at Thapar Institute of Engineering and Technology memorable.

Last but not least, I would like to thank my family whom I dearly miss and without whose blessings none of this would have been possible. To my parents, I own thanks for their wonderful love and encouragement. I would also like to thank my brother, since he always insisted me that I should do so. I would also like to thank my close friends for their constant support.

Date: 29 June, 2018

Riya Goyal

(Riya Goyal)

Place: TIET, Patiala

ABSTRACT

Millions of people in India are influenced by vision loss. Numerous advancements now are done in the field of smart assistance for visually impaired that uses ultrasound sensors and hardware-centric devices. These implementations generally increase the cost of the device making it unattainable for visually impaired. In this paper, a cheap, robust, complete walking solution is introduced, targeted at aiding the blind and low-vision people. This system is standalone and doesn't require any network access to provide assistance which in turn makes it less complex and economical. It is built by focusing on network coverage issues in the rural regions of the country. The framework is fabricated focusing on the plight of visually challenged people in India and provides specialized functionalities including object detection, face recognition and OCR coupled with audio feedback. The OCR part of the framework is built to recognize Indian regional languages like Gujarati, Hindi, Bengali etc. along with English.

The proposed device utilizes low-cost equipment and provides complete assistance to the visually challenged people focusing on the software part of the system. The facial recognition, object detection and object character recognition together help in replacing the imprecise use of traditional methods like white cane and guide dogs. With the enhancement of the applications in the field of machine vision, the scope of providing aid through a camera is endless. This framework provides precise detection resulting in a fully automated and highly accurate guiding assistant.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Blindness is one of the major disabilities affecting 39 million people around the globe. India is one of the largest countries that currently have around 12 million visually impaired people. The visual impairment [1] causes hindrance in various daily activities which creates a constant strive for the creation of assistive devices [2]. One of the difficult daily activities is recognizing people and objects around them. These problems inflict various walking hazards making them dependent on another person. The knowledge of surroundings aide those in inflecting self-confidence among visually challenged making them more self-reliable.

In our daily life, people knowingly or unknowingly participate in the social interactions in many forms such as walking, nodding, reading, writing, blinking. For visually impaired people vision plays an important role for the social interactions in any gathering with the sighted people. Researches have shown that it is not important that the interaction between two people is done only verbally, but also it can be done non-verbally. The major part of the non-verbal communications constitutes the facial expressions in comparison to the body gestures. Facial gestures are becoming the testing ground for many artificial machines as they are easy for humans to understand. Assistive devices having artificial intelligence in it are the good source of social interaction among the visually impaired people. The word “Assistive Device” refers to the one or other kind of assistance a user needs. These devices have a large scope in terms of the usage in their daily life with a major focus on user safety and well-being.

The traditional assistive aids like guide dogs and walking cane require steep learning curves and are prone to numerous limitations like the range of motion. These devices limit the information to perspective and audio. With vast improvements in the field of computer vision and processing power led to the creation of a different type of electronic traveling aids for enhancing mobility for visually impaired. Most of these enhancements are hardware-centric that increases the cost of the device making them unattainable.

An ideal walking assistant should be easily wearable, portable, lightweight and low cost so that, it could reach the majority of the blind population that is residing in the country. Since the user feedback system should be kept as real-time as possible, complex state of the art image recognition algorithms might not be viable as a relevant approach. Instead, the complexity of

these recognition algorithms should be low to permit high frame rate on less powerful hardware platforms. Generally, the mean offline social network of a visually disabled person ranges from 10 to 30 persons depending on their age [3, 4] and the relevant object detection requirement ranges from 15 to 20 objects which imply that the recognition algorithms should be trained and tested accordingly. The system should interact by providing a precise and limited feedback to the visually impaired user without overwhelming his auditory senses.

In the light of aforementioned requirements, this research presents a complete visual solution aimed at aiding visually impaired people. The system is integrated with multiple modules like face detection, object detection and object character recognition that uses highly trained machine learning models to provide high accuracy predictions.

1.1 Contribution

The system proposed in this thesis provides a one-stop solution for multiple problems faced by visually impaired. The real-time object detection system incorporated in the framework uses Single Shot Detectors (SSD) integrated with MobileNets for fast and reliable predictions. The output from object detection system is then evaluated for the presence of objects like person or text. The presence of a person in a scene activates the face detection system which further verifies whether a person is present or not using Haar cascade. If there is a person facing the camera then that person is matched with the database to discover its identity.

The two-tier architecture for face recognition helps system in differentiating an actual person from a picture or a person present in a photo frame or poster which results in making the system more practical in real-world applications. If the text is identified in the scene it triggers the object character recognition part and outputs the text recognized in the area where the text was identified. The whole system works seamlessly increasing each other reliability providing automated accurate audio feedback to the user. The OCR module built in the framework supporting many languages like Hindi, Bengali etc. that empowers the user in increasing its resource gathering areas which as of now is only limited to Braille resources.

1.2 Thesis Outline

This thesis has been divided into seven chapters. The chapter 1 introduces the subject area of the thesis research and provides an overview of the motivation and contribution in this work. Chapter 2 covers a detailed literature review of the different methodologies followed till date regarding the Object Detection, Face Recognition System and Object Character Recognition. It also explains some of the assistive devices made and used in the previous years. Chapter 3 defines the background of the system which explains the different models of object detection and face recognition in detail. Chapter 4 explains the problem statement and the objectives of the proposed approach. Chapter 5 gives an insight into the methodology followed in this research work. It includes data balancing, Preprocessing, and model building. Chapter 6 provides and discusses the experimentation results. The conclusion of the work done in this research is given in chapter 7 with the recommendations for the future work.

Chapter Summary

This chapter gives an overview of what to expect in the thesis. It gives the actual statistics of visually impaired people in India and in the world. This section describes the problems faced by them in their day to day life or in social interaction. Also, it discusses the motivation behind this thesis and the contribution to make a complete assistive system for visually impaired people. In the last it gives the thesis outline by briefly summarizing the chapters.

CHAPTER 2

RELATED WORK

Several approaches have been performed for intelligent detection and evasion of objects. Most of these systems are made up of the combination of hardware and software devices such as RFID tags, GPS, Sonar technologies etc. and are found to be precise and efficient. Some of the researches done in the field of object detection, face recognition and object character recognition are explained below.

2.1 Object Detection

White cane [5, 6] is a standout amongst the most mainstream and least complex instruments used for identification of objects. The source of its notoriety is the low cost and the portability. It is targeted at identifying any obstruction like potholes, steps, dividers etc. and so forth inside a restricted scope of the client. It is observed to be proficient in the identification of obstacles up to knee level. One of the real disadvantages of utilizing white stick is the range constraints which is restricted to one or two feet of the user depending upon the height of the stick. Besides, their ineptitudes to perceive objects like motion vehicles raised stages and so forth are a noteworthy risk to their lives. Guide dogs are another alternative for outwardly impeded. These dogs are exceptionally skilled but lack in recognition of conceivable impedance at the head level of the user. Also, the use of guide dogs requires huge adjustments in normal routine which incorporates the cost of its sustenance and nourishment.

An inventive approach was carried out by Roshni et al. [7] for furnishing an indoor positioning system with the assistance of SONAR technology. The user's position within a building or a house is identified by placing the ultrasonic units on the rooftop at standard interims. This framework was efficient in recognition of indoor objects without being impacted by any adjustment in nature. The primary disadvantage lies in the fact that this system is only compatible with the indoor surroundings.

An approach for providing indoor and outdoor locationing was performed by Nandhini et al.[8] using RFID tags for indoor positioning and GPS for outdoors. The framework [8] was

incorporated with GSM module for reaching a known person in case of emergency. The utilization of GPS [10] module implanted in the white cane for outdoor navigation decreased the cost of setting different RFID [9] labels. The main advantage of this system was its flexibility in outdoor as well as indoor applications. The utilization of RFID [11-13] tags for indoor environments increased the cost of the system making it unattainable [14-16].

Simoes et al. [17] performed a research for the creation of wearable indoor routing framework. The locationing of the user was performed with the assistance of visual markers and ultrasonic deterrents. The visual markers used in this approach helped in distinguishing distinct points in the environment. Additionally, data gathered from different sensors progressively increased the perception of the user. Wearable glasses like device integrated with multiple sensors like gyroscope, ultrasonic, accelerometer and RGB camera were used for gathering information about the environment [18]. This device helped the visually challenged client to explore its surrounding environments uninhibitedly. The main disadvantage of this system was its outdoor application constraints and the additional cost of sensors.

The Guide Cane is another novel device designed by Shoval, S et al. [19] which helps visually impaired people to quickly navigate any walking hazards and obstacles. Initially, the user pushes the guide cane in forward direction to start the operation. After that the ultra-sonic sensor identifies objects in the path of the user and directs the user towards suitable direction. The cane uses vibration in the handle which is felt by the user and help in guiding without hindering senses of user.

The Guide cane uses passive wheels to bypass its heavy weight and make the experience of the user more comfortable. The encoders embedded in the wheels helps in relative motion of the cane. The Built-in computers help the servomotor to steer the wheels of the cane in both left and right direction. The Guide cane is embedded with 10 ultrasonic sensors to detect the objects. To specify the desired direction of motion, a mini-joystick is embedded at the handle of the cane.

As discussed above, most of the approaches focus on hardware devices like RFID tags and Ultrasonic sensors to provide the feedback to the user. These approaches are not economical and require regular hardware maintenance. Thus, there was a need to create a software-centric visual assistant that provides high accuracy with minimum hardware without relying on any network connections.

2.2 Face Recognition System

An input image is given to the face detection program which is used for locating facial features by using various algorithms in the given environment [20, 21]. Face recognition method is generally used for the purpose of verification or identification. The process of identification performs facial matching of the person with the internal database containing facial data of all the known persons to confirm the identity of a person. Whereas verification is the process of confirming the identity of the person by comparing its facial data with the data of the person he claims to be.

The face recognition is applied to multiple conditions which are based on upcoming technologies, facial features and different algorithms [22, 23]. Some of those emerging methods are infra-red [24] and 3-D face recognition [25]. Both the object and face recognition are the major requirements in the field of providing assistance to the visually challenged. Many types of research have contributed to this domain but still, a user-friendly, cost-efficient and highly precise framework is required [26-28].

Verma et al. [29] introduced an image information system for acknowledging known faces. The higher resolution images used radial basis function for the administration of small training sets to balance the weight. Genetical algorithms are used in this framework for the large set of the dataset. This algorithm helps in diminishing the search and recognition time using RBF neural system. This system gives precise outdoor navigation in crowded places. The complex RBF based function constraints the system for its wide-scale application.

Kumar et al. [30] used the SIFT algorithm for the fulfillment of the basic needs of the visually impaired and also making them independent of others. Major problem encounter in their day to day life is to recognize the object and the person. The SIFT algorithm [30] was performed in the MATLAB to recover the images of objects which in turn enhance the blind's ability. SIFT algorithm includes the discovery of key points, relegating an introduction to the key points and create filter highlights. It helps the visually impaired to walk independently and enhances the portability.

Another verification approach is proposed by Raghavendra et al. [31] which comprises of both the face and speech module. The face module consists of PCA (Principal Component Analysis) algorithm for feature extraction [32, 33]. The speech module uses the speaker verification independent of text and to do the feature extraction they have used cepstral coefficients and for the opinion generator Gaussian mixture model is used. Moreover, Pong et al. [34] used the feature fusion having multi-resolution for the face recognition. To improve the data stored in the extracted features this system collects information from the face images at different resolutions. All features extracted were dependent on each other so the genetic algorithm was used in this to ensemble all the features into one vector so that it can give the favorable results.

2.3 Object Character Recognition

To recognize the words Goel et al. [35] examine the normal black and white font renderings by using whole-word subimage features. Also, Max et al. [36] performed the multiway classification of possible words from the entire dictionary to recognize the image having a word. This image is given as an input to the CNN. Although, the whole-word based recognition method outperforms but still it is dependent on lexicon system. It is necessary to ensure that the content taken from any manuscript must not be confidential.

Wang et al. [37] used the sliding window method in his research. This method is used to collect the region having character, classifies the character and after that combines the word by analyzing the classification score and the position. The author has integrated these sliding windows and applied the CNN classifier for character recognition and then used various techniques to combine the words namely beam search and Non-Maximum Suppression (NMS).

The sliding window is also adopted by Mishra et al. [38] with CRF graph for recognition of words. In this method, the classification of words involves a huge amount of training data and it is done by firstly estimating the position and size of the characters. Similarly, He et al. [39] and Shi et al. [40] executed the word recognition by using RNN (Recurrent Neural Networks) and CNN classifier.

So in this research, the proposed system uses specially designed classifiers that provides multiple features like face detection, object detection, and OCR through just a camera without the requirement of any additional sensors.

Chapter Summary

In this chapter, a detailed literature survey is carried out on three major areas which are; Object Detection, Face Recognition System and Object Character recognition. It discusses the research work going on in this area and the new techniques which have been used by different authors in the previous years. Also, the analysis has been done on their outcomes and their findings to better understand the need of the time and the research gaps left.

CHAPTER 3

BACKGROUND

3.1 Object Detection

Object Detection is a multiclass classification problem which uses a fixed size window to feed the patches of the input image to an image classifier. Each patch is sent to the classifier to predict label of the object. Some popular models of object detection are RCNN, SSD, YOLO etc.

3.1.1 Hog Features

Bill Triggs and Navneet Dalal published a groundbreaking paper in 2005 on Histogram of Oriented Gradient (HOG) features which were computationally inexpensive. Hog features from an input image are calculated by sliding window which is then fed to SVM to train the classifier. HOG features can be applied to real-world applications like pedestrian detection, face detection etc.

3.1.2 Region-based Convolutional Neural Networks (R-CNN)

Object detection is categorized as a classification problem where the accuracy of the system is dependent on a number of rightly classified objects. Development in the field of deep learning leads to the replacement of HOG classifiers with more accurate CNN classifiers. The main drawback of using CNN was that the classifiers needed to be run on a large number of patches which affected its real-time applicability.

To solve the problem R-CNN was developed which used selective search to reduce the number of patches that are fed to the model. Local cues such as texture, color, intensity assist Selective search to generate every possible object location. After applying Selective search, the reduced patches can feed to the CNN to train the classifier.

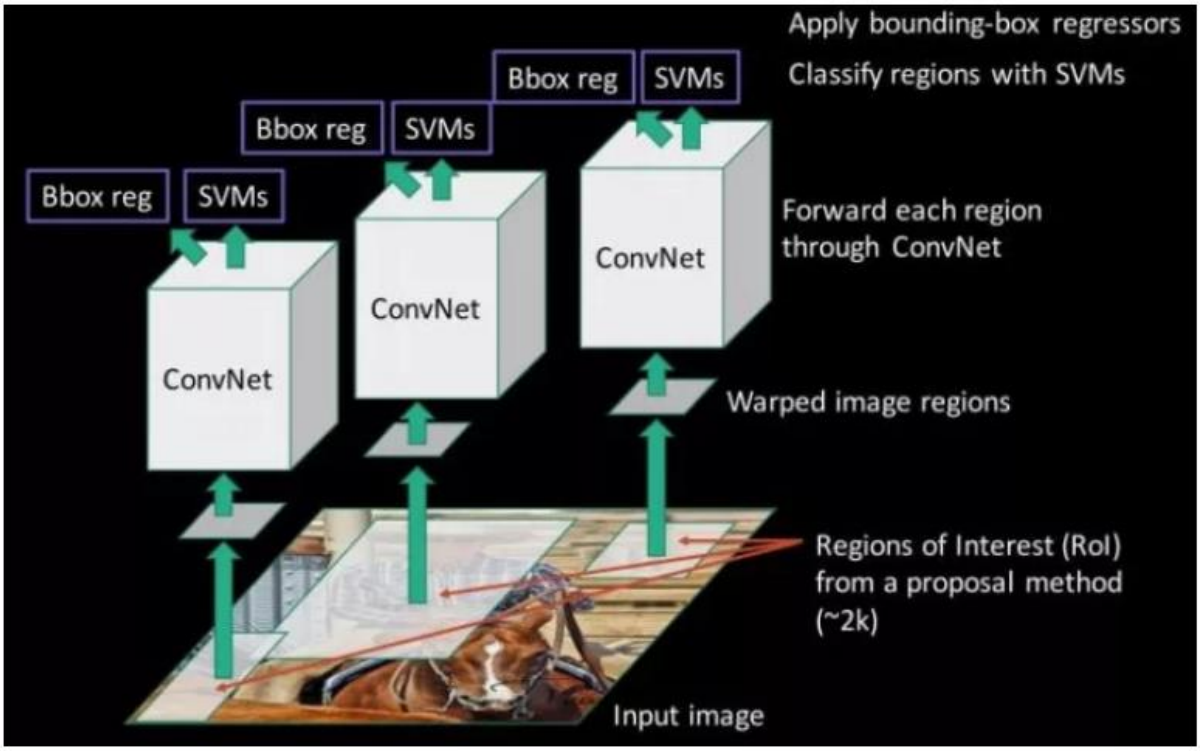


Fig.3.1.Working of RCNN [41]

3.1.3 Spatial Pyramid Pooling (SPP-net)

R-CNN despite the use of Selective search is very slow and cannot be used for real-time applications. SPP-Net tried to fix the problem of R-CNN by calculating the representation of CNN for the entire input image only one time which is further used to calculate a representation of CNN for every patch generated by Selective search. This operation can be performed by applying pooling on a section of feature map generated by last convolutional layer which corresponds to that region.

A greater challenge towards building SPP-net is that the fully connected layers of CNN required fixed size input. This problem was addressed by SPP which applied spatial pooling rather than max-pooling on the output of last convolution layer. SPP layer uses bins to divide arbitrary size region into constant size patches and the max pool is applied to the bins to generate output. Due to a constant number of bins produced every time, a constant size vector is generated as shown in Fig 3.2.

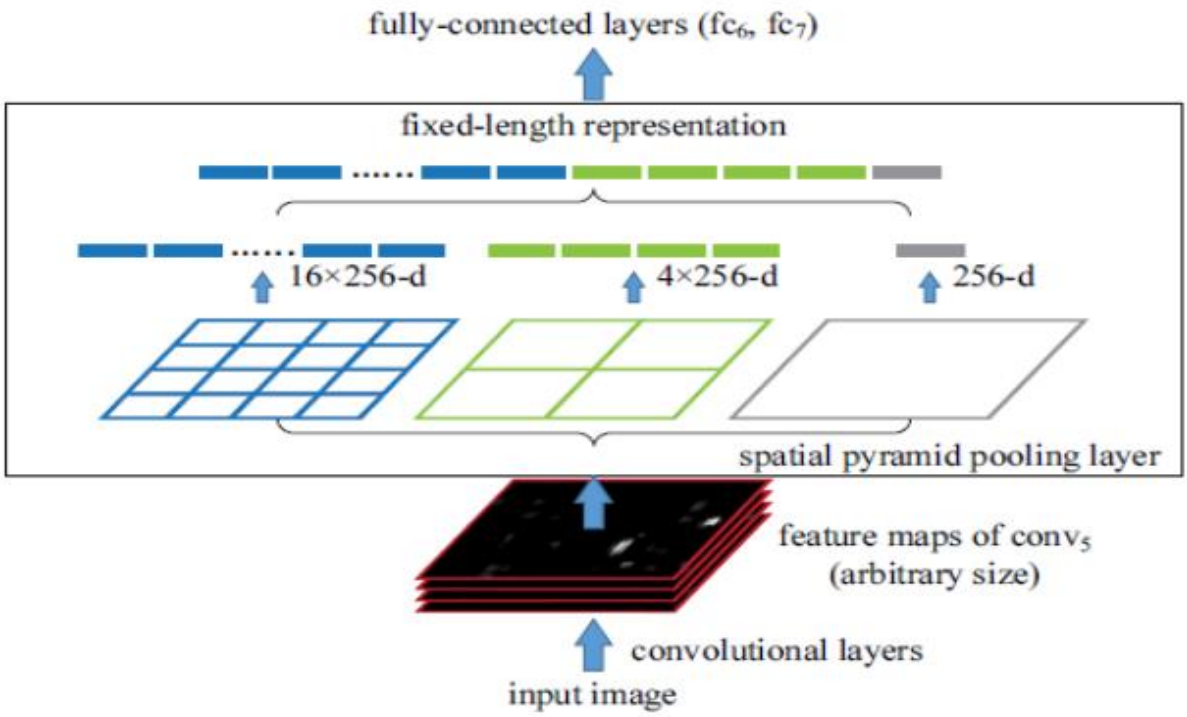


Fig.3.2. Working of Spacial Pyramid Pooling (SPP) [42]

The major drawback of SPP-net was that performing backpropagation was not trivial through spatial pooling layers. Thus the model was only able to fine tune the fully connected layers.

3.1.4 Fast R-CNN

Fast RCNN took inspiration from SPP-net and RCNN to fix the main problem in SPP-net i.e end to end training of network. Gradients are propagated using a simple back-propagation calculation that uses overlapped pooling regions. This helps gradients to pump from multiple regions of the cell. Fast R-CNN also introduced bounded box regression in the training phase of the neural network. The multitask objective of Fast R-CNN helps in simultaneous training of network for classification and localization. This reduced the overall training time and boosts the accuracy of the network.

3.1.5 Faster R-CNN

Selective Search was the most time-consuming phase in Fast R-CNN. Faster R-CNN addressed the issues of fast R-CNN by replacing Selective Search with Region Proposal Network. This network generates a region of Interests using very small convolution network as shown in Fig 3.3.

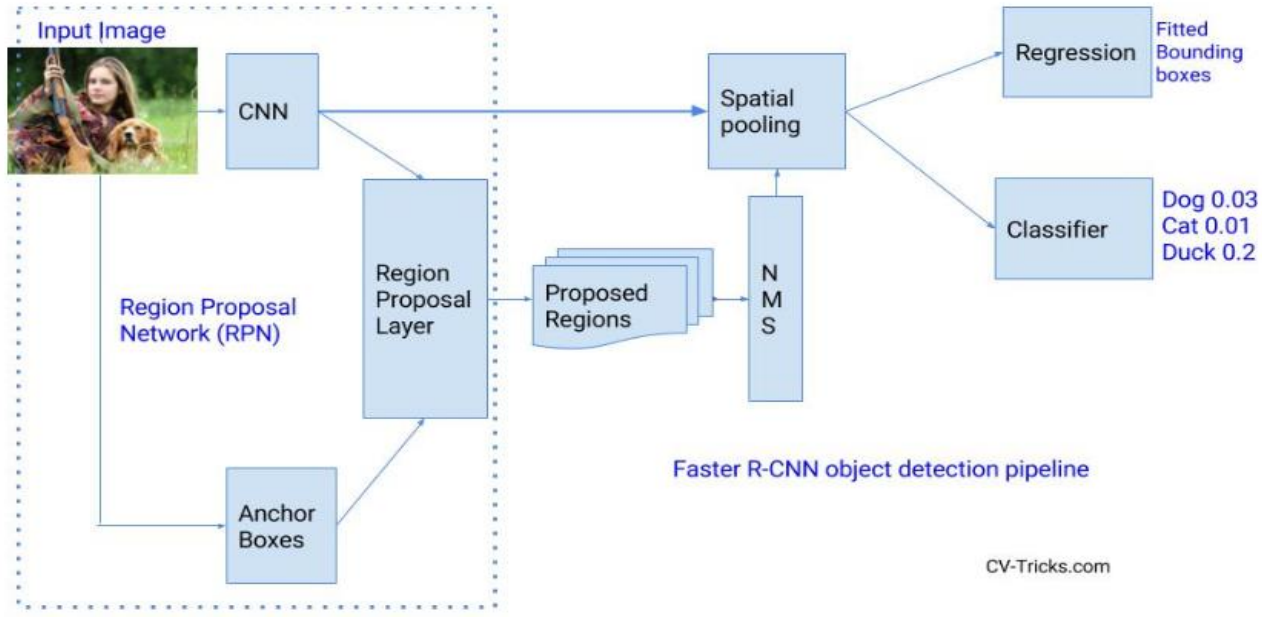


Fig.3.3.Working of Faster RCNN [42]

The idea of anchor boxes introduced by Faster R-CNN helped in handling the variations in the scale of objects and aspect ratio. Three types of anchor boxes each for different scale namely 512 x 512, 256 x 256, 128 x 128 were used at each location. Similarly, three aspect ratios were used 1:1, 2:1 and 1:2. Each location in total has 9 anchor boxes which are used to predict the probability of them being foreground or background using RPN. Bounded box regression is applied at each location to improve the anchor boxes. Probabilities and bounded boxes of different sizes are returned by RPN. Variation in size of the bounded boxes is passed by applying Spatial Pooling. Faster R-CNN has 10 times faster prediction speed than Fast R-CNN with almost similar accuracy. A quick comparison between different versions of R-CNN is shown in Table 3.1

Table 3.1. Comparison between the three R-CNN models

	R-CNN	Fast R-CNN	Faster R-CNN
Test Time per Image	50 Seconds	2 Seconds	0.2 Seconds
Speed Up	1x	25x	250x

3.1.6. YOLO (You only Look Once)

YOLO portraits object detection as a simple regression problem that takes an input image and learns its bounded box coordinates and class probabilities. YOLO partitions each input image into a grid of size SxS and every grid gives a prediction of N bounded boxes and confidence. Confidence depicts whether an object is actually contained in a bounded box or not. Thus a total number of boxes predicted is SxSxN. Most of the predicted boxes have very low confidence, so by applying a threshold of 30%, we can remove most of them as shown in Fig 3.4.

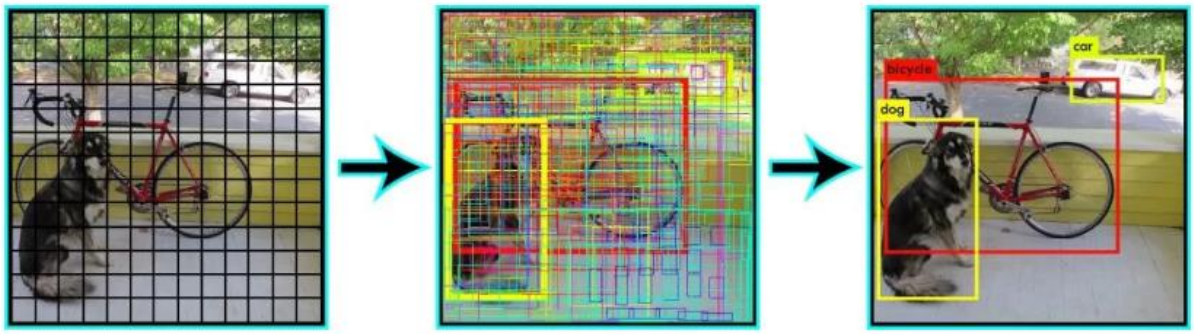


Fig.3.4. Working of YOLO model [42]

Yolo uses CNN only once and hence can be applied to real-world applications. One main difference between YOLO and rest of object detection models is that it sees complete image rather than generated small regions. This contextual information aids in minimizing false positives. The major drawback of using YOLO is that it predicts only one class in one grid and hence finds it difficult to identify small objects.

3.1.7. Single Shot Detector (SSD)

SSD provides a balance between accuracy and speed. It calculates the feature map by applying CNN on input image one time. After that, a convolutional kernel of size 3x3 is applied on feature map to predict classification probability and bounded boxes. SSD learns the off-set by applying anchor boxes at different aspect ratios and predicts bounded boxes after multiple convolutional layers. Due to a different scale of operation of each convolutional layer, SSD is able to predict objects of various sizes.

3.2 Face Detection

Paul and Michael proposed an effective method of object detection that uses Haar-like feature based cascade classifier [6]. Firstly training of Haar cascade is done by using positive and negative examples scaled in the same size. After training the classifier, the input image is portioned into rectangular regions to search for the presence of human faces. If the classifier identifies a human face in the region of interest then it gives '1' as output otherwise '0'. In order to search faces of different size [7], a resizable search window is used. Haar cascade became exceedingly popular fro face detection as it can be easily developed by using Open CV classifier and has a high accuracy rate.

3.2.1 Eigenfaces

It was proposed by Pentland and Turk[8] and was based on Principle Component Analysis(PCA). Eigenfaces help in dimension reduction of image matrix. It uses linear transformation method to map the n-dimensional space of sample image to m-dimensional feature space where m < n. Total scatter matrix (ST) is defined as:

$$S_T = \sum_{i=1}^{N} (x_k - \mu) \cdot (x_k - \mu)^T \quad (1)$$

Here μ represents mean image of the sample set, N is a total number of images in the sample, and x_k is the columns concatenated image in a vector.

3.2.2 Fisherfaces

It solves the light condition sensitivity in face recognition [9] problem by using a linear method. To reduce the dimensionality of the image, it makes use of the class-specific linear method and further classifiers were used to reduce feature space. In order to increase the reliability of classification, shaping of scatters was performed by the class-specific method. The between-class scatter matrix defined as in (2).

$$S_b = \sum_{i=1}^{C} N_i (\mu_i - \mu) \cdot (\mu_i - \mu)^T \tag{2}$$

The within-class scatter matrix defined as in (3)

$$S_w = \sum_{i=1}^{C} \sum_{x_k \in X_i} N_i (\mu_i - \mu) \cdot (\mu_i - \mu)^T \tag{3}$$

In equation 2 and 3, μ_i represents the mean image of X_i class, μ represents the mean image of all classes. Total number of classes is represented as C and N_i is the total number of images of class X_i. S_w is kept a minimum while S_b is made maximum fr performing classification [10].

3.2.3 Local Binary Patterns (LBP)

Local Binary Pattern was initially introduced as a method of texture description, that divides the image into local regions. For every region, labeling of each pixel was done using decimal value. These decimal values are then grouped into a histogram and computations are performed for histogram similarity to execute classification [11]. Ahonen [12] introduced LBP based face description in which image containing face is split into local regions and the extraction of LBP description is performed on them independently. Division of each region is dived into the neighborhood of 3x3 pixels where the center pixel represents the threshold and result is obtained in for of binary number. If the value obtained is greater than threshold than binary value is “1” otherwise “0”. After the process an eight digit binary is produced, this is again converted into decimal and represent as LBP code of the region [10].

Chapter Summary

This chapter is mainly divided into two subsections namely Object detection and Face Detection. In the first subsection, the detailed working of seven different object detection models is discussed with their pros and cons. In the second subsection, the working of its three models is explained along with their mathematical formula. Moreover, analysis is done over how each successive model is an improvement over the previously discussed model.

CHAPTER 4

PROBLEM STATEMENT

This research addresses the issues faced by visually impaired by proposing an assistive wearable device. Most of the assistive devices available in the market use RFID, Sonar and IR sensors etc. However, each of them possesses certain drawbacks moreover no device is with complete assistance to blind which includes the face, object and object character recognition. Use of high-end sensor increased the cost of the assistive device making it unattainable for poor. With the advancement in the field of deep learning and image classification, most of the work which required multiple sensors can now be replicated with the help of a camera. The major challenge regarding this approach was to create a real-time system which has high frame rate and accuracy. The System should have no network dependencies so that steady performance is achieved irrespective of the location and area of usage.

4.1 Aims and Objectives

The objectives of the current research work can be listed as:

1. To study, analyze and explore the already existing assistive devices for visually impaired.
2. To compare existing machine learning models for the task of Object, Face and Object Character Recognition and select the appropriate one for the given task.
3. To propose cheap and reliable complete assistive solution for visually impaired using auditory feedback.
4. To introduce such a device that can fulfill the basic requirements of a visually challenged person so that he or she can participate independently in any social interactions.
5. To test and validate the proposed assistive device on custom datasets and gather user feedback.

CHAPTER 5

PROPOSED ASSISTIVE DEVICE FOR VISUALLY IMPAIRED

The architecture of the proposed system as shown in Fig. 5.1 consists of 3 main modules namely object detection, face recognition and Object Character Recognition coupled with text to speech module for audio feedback.

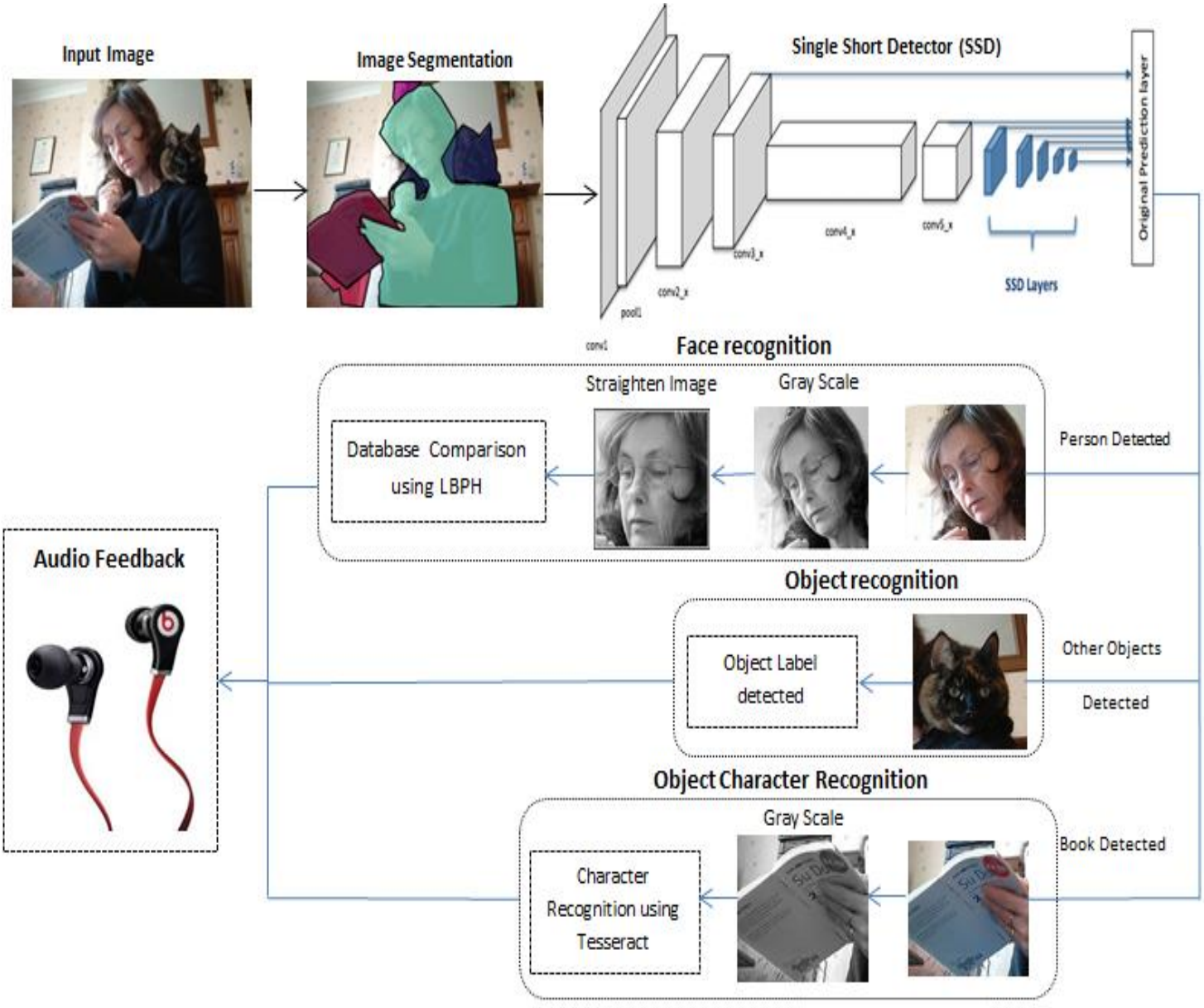


Fig 5.1.The architecture of the Proposed System

Initially, a frame is captured from a video stream and is passed to the object detection module. The custom model is used for the detection of an object. This custom model is trained by the segmented image for further recognition of objects. The model is based on the Single Shot detector which is an extension of CNN that uses multiple convolutional layers to detect objects in real time. After the object is detected it is further evaluated for the presence of a person.

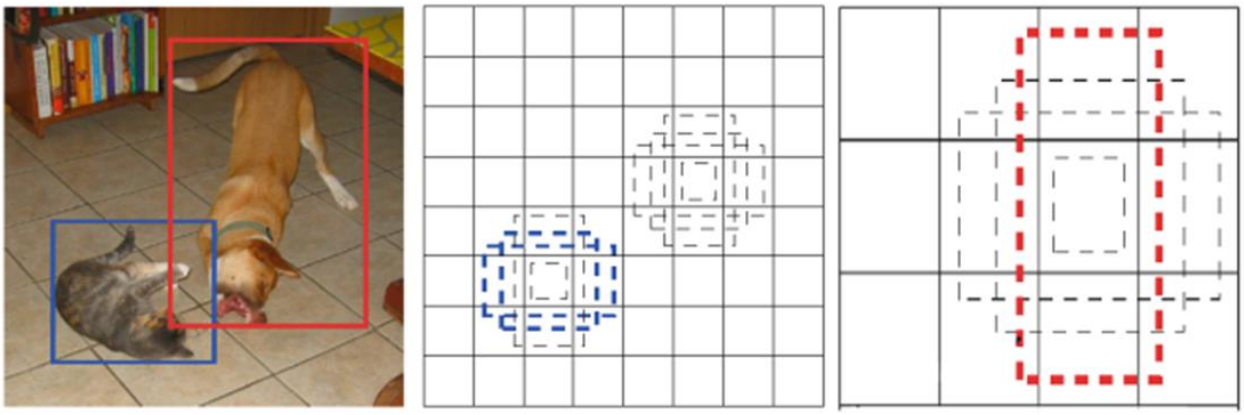
If a person is identified in the frame the system triggers the face recognition module. This module initially identifies the presence of the facial coordinates in the ROI, i.e., the region where the person is present in the frame. If the facial coordinates are present then they are further extracted from the frame to apply the multiple preprocessing techniques like grayscale and image straightening. These preprocessing steps increase the detection rate of the system. The preprocessed image is then compared with the database using trained face recognition model to recognize the label.

If the text is identified in the frame then OCR is activated. OCR uses Tesseract model that converts the text present in the image into a string. The text which is converted is further sent to the text to speech module for initiating an audio feedback to the user. Rest all the object labels categorized into “Other Objects” category are directly sent to audio feedback unit to create an auditory response. Each module of the above architecture assists in working with other modules creating a reliable low-cost device which performs with high accuracy using minimal hardware. The detailed description about the different modules has been explained in next subsections.

5.1 Object Detection

Three most common deep learning-based models for the purpose of object detection are Faster R-CNN's, You Only Look Once (YOLO) and Single Shot Detectors (SSDs). The Faster RCNN is an extension of RCNN that mainly focuses on the Speeding Up Region Proposal. RCNN uses selective search process for creating the bounding boxes, or region proposals. This process of selective search makes RCNN extremely slow and unusable for real-time applications. The advantage of Faster R-CNN is that it bypasses selective search and uses features gathered in the forward pass of CNN for generating region proposals. This model accomplished high accuracy, however, it was difficult to comprehend and prepare which influences its wide-scale

implementation. Additionally, even the fastest implementation of R-CNNs, i.e., “faster R-CNN” could just accomplish up to 7FPS (frame per sec) which influences their real-time usage. This effect in speed was addressed by YOLO [43] model which is the fastest object detection classifier. The model looks at the whole image at test time so its predictions are informed by the global context in the image. It also makes predictions with a single network evaluation unlike systems like R-CNN making it 100 times faster. With appropriate, hardware YOLO model can detect objects at up to 155FPS.



(a) (b) (c)

Fig.5.2. SSD framework (a) Image with GT boxes (b) 8*8 feature map (c) 4*4 feature map [42]

The main drawback of YOLO model is that high speed was achieved at the expense of accuracy. SSD [44] developed by Google gave a decent harmony amongst accuracy and detection speed providing a simpler and straightforward approach than complex R-CNNs [45] and YOLO.

In our proposed system, the SSD model is integrated with MobileNets for faster and accurate detection of objects. SSD is an extension of feedforward CNN architecture. It recognizes objects in an image by using fixed size bounded boxes as shown in Fig. 5.2. Each bounding box is assigned a score indicating the presence of object followed by suppression to produce output.

The initial layers of network use standard architecture for image classification task followed by auxiliary structure to produce results by using the Multi-scale feature maps for detection. The convolutional layers are integrated at the later part of the base network and progressively decrease the size. These layers help in detection of objects at different scales. Every convolutional layer uses convolutional filters to produce predictions on a fixed window

size. Each bounding box is associated with its feature map cell. Each cell predicts the presence of an object by returning the offset value and per class score corresponding to the box shape. Four offsets and 'c' class scores are computed from each box out of' given conditions which result in(c + 4)*d convolutional filters. These filters are applied at every location of the feature map of dimensions a*b generating outputs as in (4).

$$(c + 4) * d * a * b \qquad (4)$$

Fig. 5.2 illustrates the framework of SSD. Default boxes implemented in SSDs' are similar to the anchor boxes used in Faster R-CNN but are applied to multiple resolution feature maps. Varying the shape of default boxes aids in identifying the possible shape of output box. Real-time object detection models require special attention towards the size of network architecture.

Object detection architectures like SSD can be in the order of 200-500 MB and thus can't be implemented in resource-constrained devices. The proposed system uses a low processing power device for building a walking assistant for visually impaired. Thus, special steps need to be taken to reduce the size of SSD to make it more power efficient. MobileNets developed by Google researchers become exceedingly popular among the devices with low processing power. The main difference that separates MobileNets from other convolutional neural networks is their usage of depth-wise separable convolution. Depth-wise separable convolution is implemented on the idea of splitting the convolution into two stages, i.e., depth-wise convolutions filter size 3*3 and followed by a pointwise convolution of 1*1 filter size. This setup reduces network parameters on the tradeoff of accuracy. These nets are not as accurate as their large counterparts but their low resource consumption increases their applicability.

Combination of Single Shot Detectors (SSD) architecture and MobileNets framework creates an efficient and fast deep learning system for the purpose of object detection. Initially, images from COCO (Fig. 5.3) dataset are used for training MobileNetsSSD. After that network is fine-tuned on PASCAL VOC reaching 72.7% mean average precision. The classes used to train our model include *airplanes, bicycles, birds, boats, bottles, buses, cars, cats, chairs, cows, dining tables, dogs, horses, motorbikes, people, potted plants, sheep, sofas, trains,* and *TV monitors*.

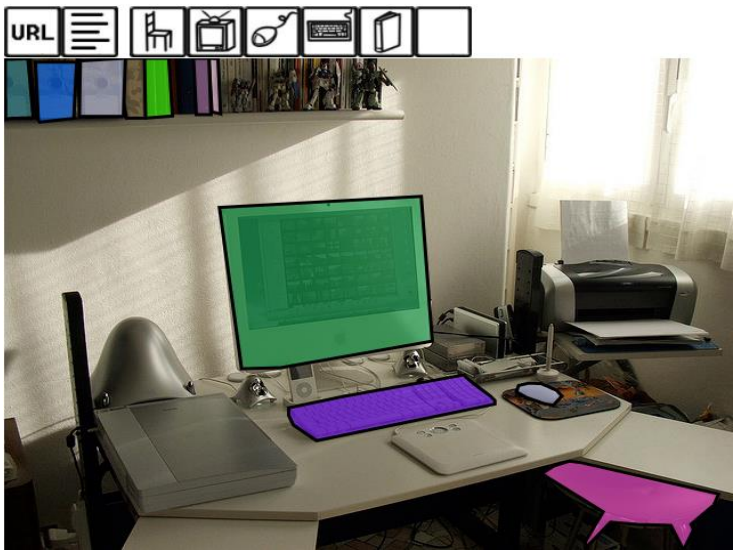


Fig.5.3. Image from COCO dataset showing object segmentation

5.2 Face Recognition

Networks trained for Face recognition are mostly tested on benchmark datasets as mentioned in related work section. Publically available face datasets include AT&T dataset, XM2VTS Database, the Oulu Physics Database, the Yale Face Database, Purdue AR Database, Illumination and Expression Database, the CMU Pose, the MIT Database and the FERET Database. Providing an efficient and robust face recognition algorithm requires the algorithm to differentiate between minute changes in the images and disregarding additional image content of surroundings and illuminations. The success of a face recognition algorithm is highly dependent on the quality of data used for training. Training database with less variation in pose angle and illuminations generally complicates the differentiation of two unique faces.

Most of the databases available do not use calibrated setups for data acquisition. To remove the dependence on the publically available dataset, database FaceX is created, which contains face images with varying poses and illuminations. A subset of the created database is shown in Fig.5.4.

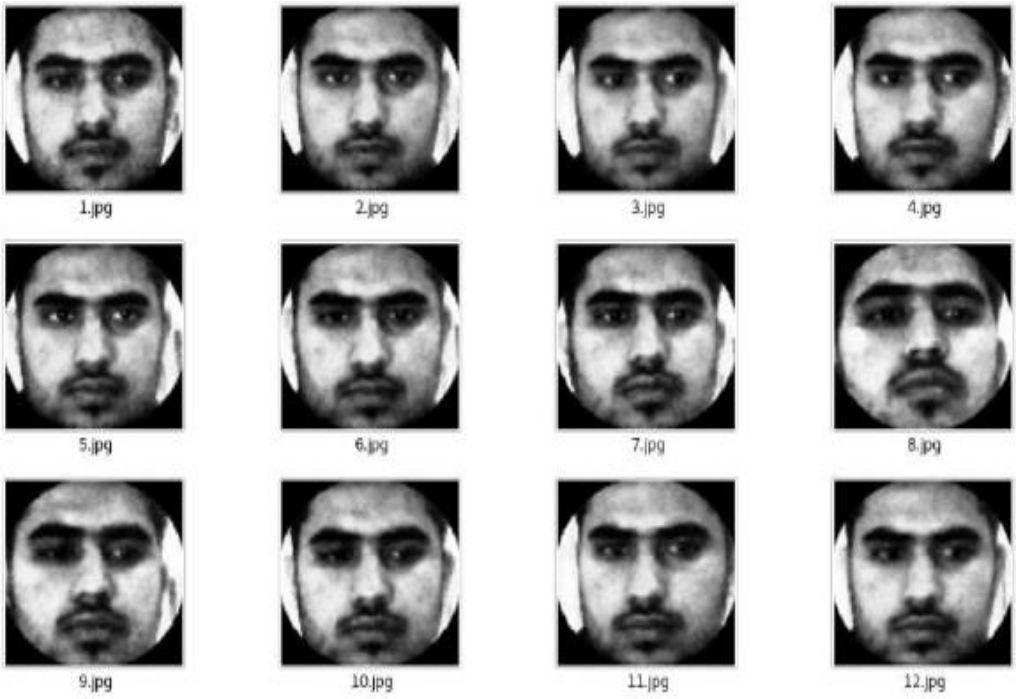


Fig.5.4. Sample frames from created dataset FaceX

Setup for data acquisition includes a video camera and rotatable spotlight with the subject seated in the center. The extraction of frames is automated on 30 users with the help of video sequence. Database for every user is created using two steps. Initially, 362 images are captured(Two frames per angle) varying the pose angles between +90 degrees and –90 degrees. After that 62 images are captured (Two frames per angle) varying the illumination angles between +90 degrees and –90 degrees.

The created database can be visualized as two dimension matrix where rows of the matrix represent unique users and columns represent variation in pose and illuminations. Face images from each frame of the database are extracted and resized at 128 pixels wide and 128 pixels high to standardize the size of each image of the dataset. Images are resized keeping their aspect ratios preserved to avoid irregularity in the facial structure. Each face image is normalized by placing eyes on the subject on 57^{th} row and mouth at 87^{th} row from the top of the image. This is done to create a similar standardized centered position for all facial features like eyes, mouth, nose etc.

This created database aids in the training of face recognition models. The initial step for applying face recognition over a set of images is to isolate facial region. For this task, an algorithm that

uses Haar cascade and AdaBoost is applied as shown in Fig. 5.5. A frame acquired video stream is divided into fixed size overlapping regions. Bank of filters is applied to these regions to identify the presence of facial coordinates. These banks of filters (Rectangle filters) use variations in intensity to identify different facial features.

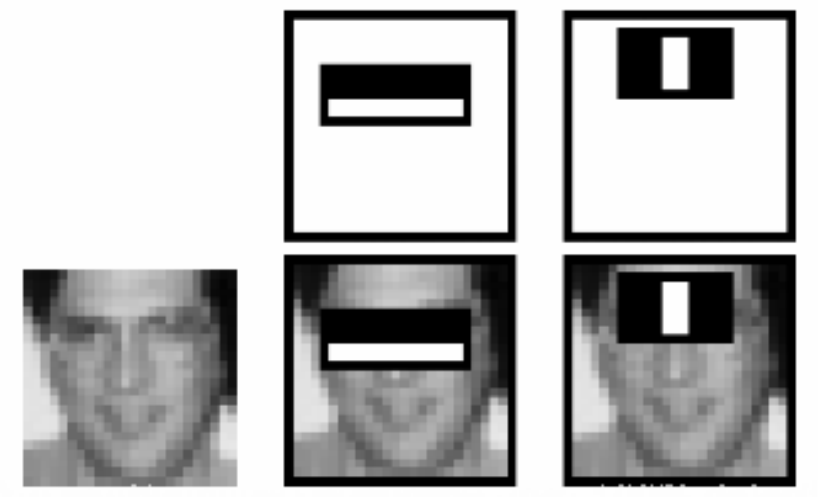


Fig 5.5.Working of Rectangle filters [46]

For instance, forehead region of the face is generally of high intensity when compared with the eye area as shown in Fig.5.5. A filter of rectangle shape used to identify such shift in intensity would be as wide as the width of the facial region. Its breadth divided into white region representing the forehead area and black region for the eye sockets. Cascade of filter banks is used to reduce the time consumed in analyzing every region of the frame.

Cascade is configured by placing fewer filters at the start which increases the false positive by decreasing detection time of the system. Large numbers of filters are placed at the end of the cascade which improves the accuracy and time for processing of the image. This two-stage detection decreases the time for rejecting frame without the presence of human face at beginning layers of the cascade. For a face to be identified, it has to pass through all cascade filters. This cascade configuration decreases the processing time making the algorithm suitable for real-time applications. Fig.5.6 shows the face detection algorithm recognizing face regions in the video stream.

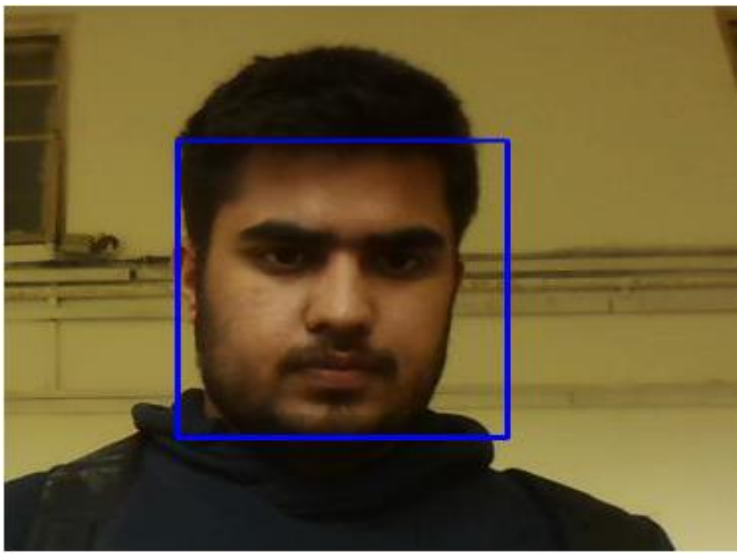


Fig.5.6. Bounded box showing the face region in the video stream

After the face is identified in the input image next step is to match the detected face with the database to recognize the person. The best approaches to identify the face of the person are namely Eigenfaces, Fisherfaces, and LBPH. Both the Eigenfaces and Fisherfaces try to reduce the High Dimensional data using a holistic approach. Eigen Face approach faces difficulty when variance originated in the image due to external sources. In this case, the maximum variance doesn't mean the information is present in these components. To preserve the discriminant information from being lost LDA is applied using Fisherface. This method works great in scenarios of the constrained environment but fails to provide the same amount of accuracy in a real-world scenario. This shows that the exact light settings cannot be assured in only 10 images of a person. The question arises when there is only one image per person, in that case, the image recognized may be horribly wrong. To get good results one at least needs 9 (+-1) images per person so fisherface method fails here. But there are other factors that affect the image like rotation, scaling etc. LBPH helps tackle these difficulties by comparing the intensity of each pixel with its neighborhood to summarize the local structure of the image. If the given pixel's intensity is equal to or greater than its neighbor pixels than it is tagged with 1 otherwise 0. So if we take 8 neighbor pixels than each pixel is denoted by 8-bit local binary pattern as 11101111. The description of the LBPH operator is shown in equation (5).

$$\text{LBPH}\ (a_x, bo_x) = \sum_{q=0}^{q-1} 2^q \text{S}(\text{jq} - \text{jx}) \qquad (5)$$

with (a_x, bo_x) as a central pixel with intensity j_q; and j_n being the intensity of the neighbor pixel. s(x) is the sign function defined in equation (6).

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases} \quad (6)$$

In these, very minute details of the image are obtained. During our evaluation, LBPH Model performs the best out of the three models so for further evaluations LBPH is selected as the base model. This model is highly robust towards grayscale monotonic transformations. Once the model is trained then the label that contains the name of the identified person is returned from the model that label is sent to the python text to speech library.

5.3. Object Character Recognition

In the framework, *Tesseract* is chosen for the purpose of detecting words from image frames. *Tesseract* is one of the best widely accepted algorithm for the purpose of implementing object character recognition. It accepts the binary image as input for detecting text using traditional pipeline architecture. The storing of component's outlines is done using component analysis. This step is considered expensive but has multiple advantages. Detection of nested, child and grandchild outlines eases the task of detecting inverse text making the task similar to the detection of black on white text.

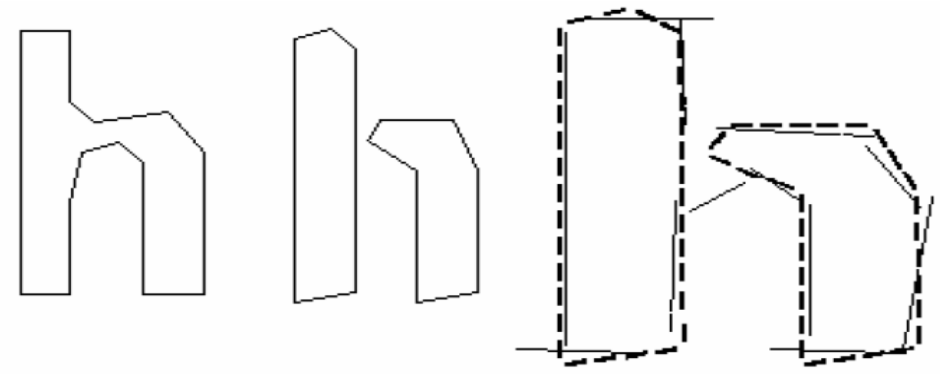


Fig.5.7. (a) Pristine 'h' (b) Feature 'h' (c) feature matched to prototype

Fig. 5.7 shows the working of Tesseract for detecting word 'h'. Outlines are clubbed together creating blobs which are further arranged into text lines and pitch text. The lines are broken

down according to character spacing to identify words. The pitch text is partitioned by character cells. The final recognition phase uses two-pass processes. The first pass tries to identify words which are then fed to the adaptive classifier for training purpose. The adapting classifier when trained recognizes text more accurately in the later part of the page. Due to late training, there is a chance that the classifier failed to detect words near the top of the page. The second pass mitigates the shortcomings of delayed training by applying the classifier again on the whole image thus identifying words which were not recognized in the first phase. Finally, all fuzzy spaces are resolved and alternate hypothesis is checked to detect smallcap text. The system is able to recognize English along with Indian regional languages like Gujarati, Bengali, Gujarati, Hindi, Marathi, Panjabi, Telugu, Sindhi, Tamil and Urdu.

5.4 Text-to-speech Converter

After the detection of objects and person, audio feedback is generated to notify the user. Python 'pyttsx' library is applied to power the audio engine converting text input to speech. This is further fed to wireless headphones placed in the user's ears. To increase the accuracy and reduce misdetection of persons, five consecutive frames are analyzed. If all frames return same output then output from face detection is considered correct and sent to the text to speech converter otherwise it is discarded.

5.5 Hardware Integration of Assistive Device

The framework introduced in this approach consists of three major components namely Raspberry Pi 5 camera, Raspberry Pi 3 Model B motherboard and Wireless Bluetooth earpiece. The Raspberry Pi 5 MP camera module is used for video stream acquisition. This camera is housed in the cane of the user. The camera has a resolution of $2592 * 1944$ pixels and can record up to 1080p videos at 30 frames per sec. The camera is attached to a ribbon cable attached to the camera serial interface (CSI) of the Motherboard.

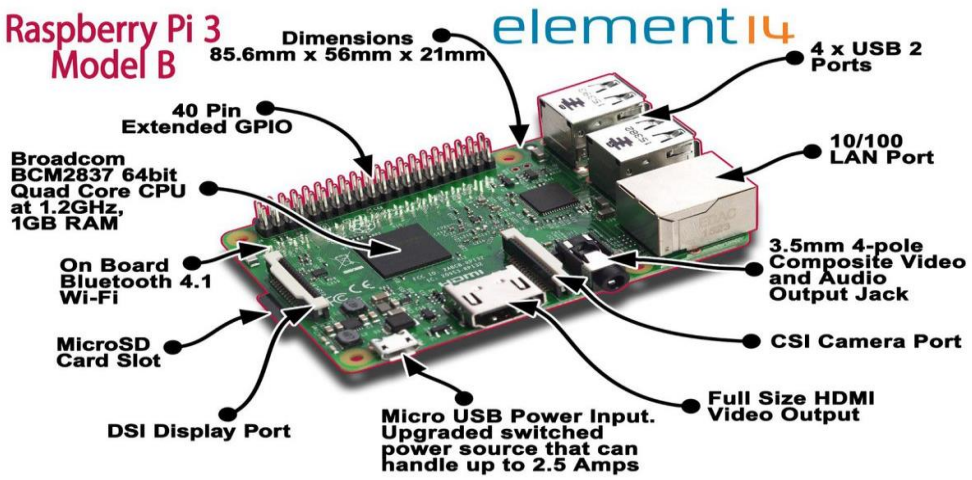


Fig.5.8. Detailed Description of Raspberry Pi 3 Model B

Fig.5.8 shows the detailed demonstration of the Raspberry Pi 3 Model B motherboard. It is used to power the whole framework. This integrated chip is built on latest Broadcom 2837 ARM v8 64 bit processor with 1 GB ram and is capable of processing at a speed of up to 1.2 GHz. The board is integrated with wireless LAN and Bluetooth 4.1 for providing wireless connectivity. This chip is housed in the cane along with the camera. The whole setup uses 5000mA power bank which provides up to 10hrs of battery backup.

(a) (b)

Fig.5.9. Hardware Integration of proposed assistive device (a) setup showing Raspberry Pi board and camera (b) Showing the case cover of both Pi board and camera.

The complete setup of the proposed assistive device is shown in Fig.5.9. Here, the camera is attached to the CSI camera port on the motherboard and the SD card in the Micro SD card slot. The motherboard and the camera are covered with its case and make this device very compact and lightweight. This is highly portable and cost-efficient as compared to the other fabricated devices made till now.

To make the board usable a suitable operating system is installed which supports all the python libraries used to build our model. There are multiple operating systems that support Raspberry Pi like Ubuntu core, Windows 10 IoT but for our system, we choose Raspbian OS which is based on Linux. The Raspbian OS comes with an auto-login feature enabled and has a very minimalistic interface which helps in smooth running of Raspberry Pi. After that, the code is placed on Raspberry Pi and set up to run in the background on every boot. A Wireless Bluetooth earpiece is connected to the Raspberry Pi. This earpiece provides audio feedback to the user without hindering his auditory perception.

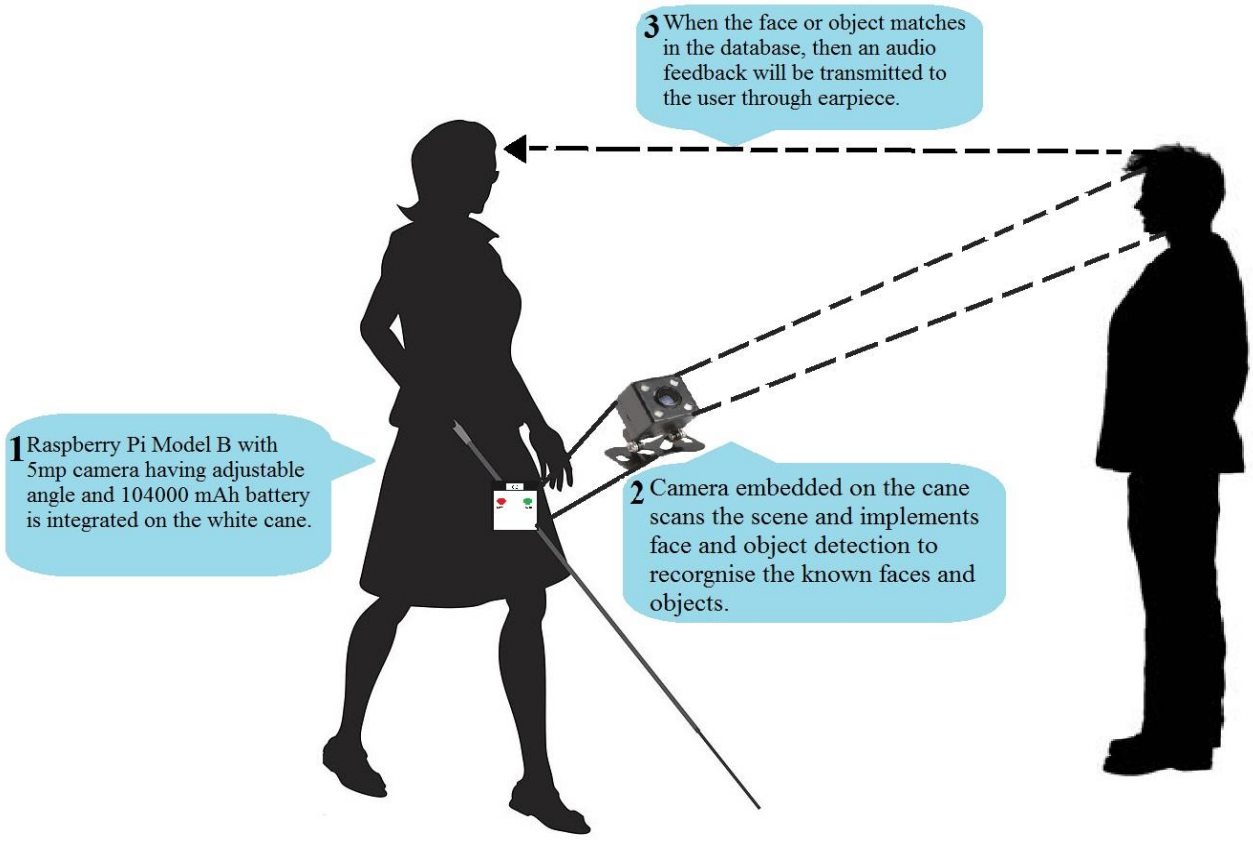


Fig.5.10. Descriptive diagram of walking assistant

Fig.5.10 shows the entire setup of the portable assistive device. Here, the device is mounted on the white cane with adjusting camera pointing in the field of view of the user. The camera continuously captures the frames and evaluates them for the presence of object, person or text. The resultant predictions are then sent to the text to speech module which provides audio feedback through wireless headphones to the user. In this way, the proposed system assists the visually impaired in the daily chores.

Chapter Summary

In this chapter, the architecture is proposed for the assistive device for visually impaired people. Each and every step is explained starting from the capturing of input image from a video stream to firstly detect the object in it by using SSD model. If the image is having an object than it will display the label of that object but if the object is person than it goes to the face recognition module. This module recognizes the facial region and after that identifies the identity of that person from the database and gives the audio feedback to the user by converting text to speech. Also, it explains the working of object character recognition module to recognize the text if present and gives its auditory output. Lastly, it explains the detailed hardware integration of a portable assistive device made for the visually challenged people.

CHAPTER 6

RESULTS AND DISCUSSION

6.1 Evaluation Metrics

In order to check the quality of the proposed system, various parameters have been used. Their efficiency and authenticity are evaluated by using the formulas and defining necessary terms as follows.

6.1.1 Accuracy

Accuracy helps in evaluating the performance of the framework. It is the measure of rightly classified instances among all the instances. The higher the accuracy the higher is the number of instances correctly classified into their classes and thus the better is the model. This can be calculated as in (7).

$$\boldsymbol{Accuracy = \frac{Rightly\ classified\ instances}{Total\ number\ of\ instances}} \quad \textbf{(7)}$$

6.1.2 mAp (mean Average precision)

Mean Average Precision (MAP) is the standard single-number measure for comparing search algorithms. Average precision (AP) is the average of precision values at all ranks where relevant documents are found. AP values are then averaged over a large set of queries as shown in (8), (9).

$$\boldsymbol{AP = \frac{1}{|R|}\sum_{i=1}^{n} Prec(i).relevance(i)} \quad \textbf{(8)}$$

$$\boldsymbol{MAP = \frac{1}{|Q|}.\sum_{Qi \in Q} AP(Qi)} \quad \textbf{(9)}$$

where,

R →Total number of relevant instances.

Prec (i) →Precession at the top of the i^{th} instance.

Relevance(i) →1 if relevant, otherwise 0.

Q → Total number of Queries.

6.1.3 Frame rate Per Sec

The frame rate is the recurrence (rate) at which consecutive images called frames show up on a display. The term applies similarly to film and PC designs, and movement catch frameworks. Frame rate may likewise be known as the frame frequency, and be communicated in Hertz.

6.2 Experimentation and Results

As the device proposed is user-centric, its success completely depends on how useful the user thinks it was. This approach distinguishes itself from the rest devices in the market by focusing on affordability and less reliance on an internet connection which makes it an effective device even in the most remote areas of the country.

The complete system is powered by Raspberry pi's latest Model B motherboard and focuses on the software part of the framework for the object, face and OCR recognition rather than integrating different sensors which elevate the cost. The implementation of python's text to speech library enables the system to provide a TalkBack feature for improving the auditory perception. The system is completely off the grid providing complete data security and reliability. The framework proposed in this approach explains the importance of face, object and character recognition in the daily life of visually impaired people. The experimentation and results detail on the proposed system has been given in the next subsections.

6.2.1 Face Recognition

After detecting a face region in the video stream, additional analysis is implemented to identify the face from the created database. Comparison of different detection algorithm in Fig.6.1 shows that LBPH outperforms other two algorithms, providing steady results with varying pose and illuminations. Testing of algorithms was done in the real-world situations. Set of 500 images were taken for 10 different people in an office situation and 5 trials were executed to verify the performance of three algorithms.

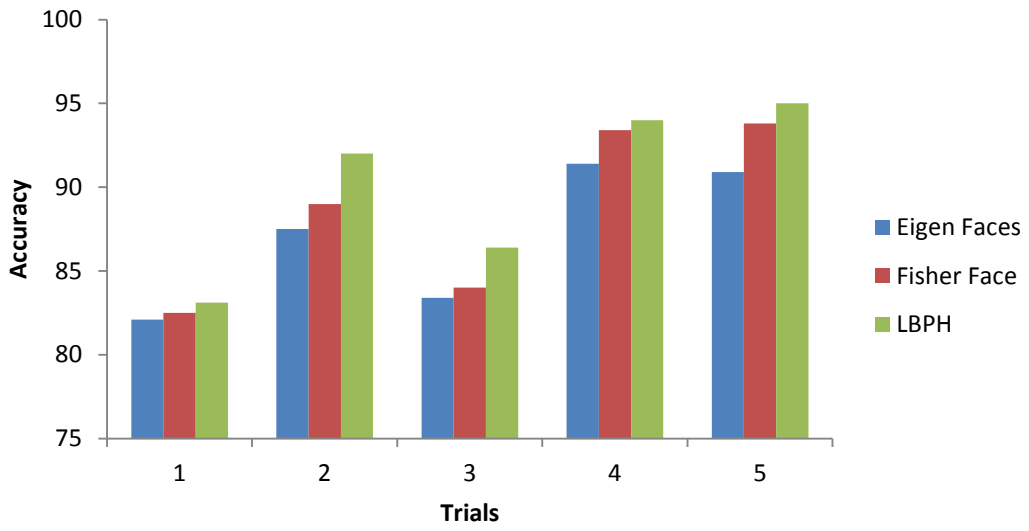


Fig.6.1. Accuracy v/s trials of three face detection algorithms

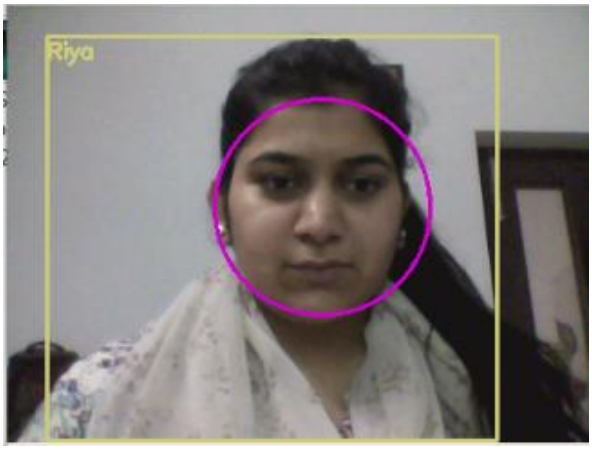


Fig.6.2. Identifying the face region with the label name

Fig.6.2 shows the face region from a video stream and matched the image with the database to recognize the identity of the person. It also shows the label if the person is known. Table 6.1 shows the rightly classified images of each face recognition algorithm in every trial. After taking the average of rightly classified images for each algorithm, it is concluded that LBPH has the highest number of rightly classified images amongst the three. Thus, LBPH algorithm is used for further experimentation of the proposed system.

Table 6.1.Comparison of accuracy of different face detection algorithms

S.No	Total number of test Images	Face Detection algorithms showing rightly classified images		
		Eigenface	**Fishersface**	**LBPH**
1.	5000	4105	4125	4155
2.	5000	4375	4450	4600
3.	5000	4170	4200	4320
4.	5000	4570	4670	4700
5.	5000	4545	4690	4750
Average	**5000**	**4353**	**4427**	**4505**

6.2.2 Object Detection

Object recognition algorithm acts as a base for the proposed system. To choose the best algorithm, a series of experimentations were performed over the most popular object detection algorithm. For evaluating the performance of the algorithm, three main parameters were selected namely mean average precision, frames per sec and number of bounded boxes created.

Each object detection algorithm was tested on PASCAL VOC 2007 benchmark dataset. Table 6.2 shows the comparison between SSD, Faster R-CNN, and YOLO. SSD method outperforms Faster R-CNN in both speed and accuracy. Although Fast YOLO can run at 155 FPS, it has lower accuracy by almost 22% mAP. To the best of our knowledge, SSD is the first real-time method to achieve above 70% mAP.

Table 6.2.Results on Pascal VOC2007 test

Method	***mAP***	**FPS**	**#Boxes**
Faster RCNN (0VGG16)	73.2	7	300
Faster RCNN (ZF)	62.1	17	300
YOLO	63.4	45	98

Fast YOLO	52.7	155	98
SSD	74.3	46	8732

Fig. 6.3 shows the graphical description of the mean average precision and the frames per second of the different object detection models. The comparison proves that SSD outperforms the other two algorithms by obtaining 74.3% mAP on 46 FPS having 8732 bounded boxes.

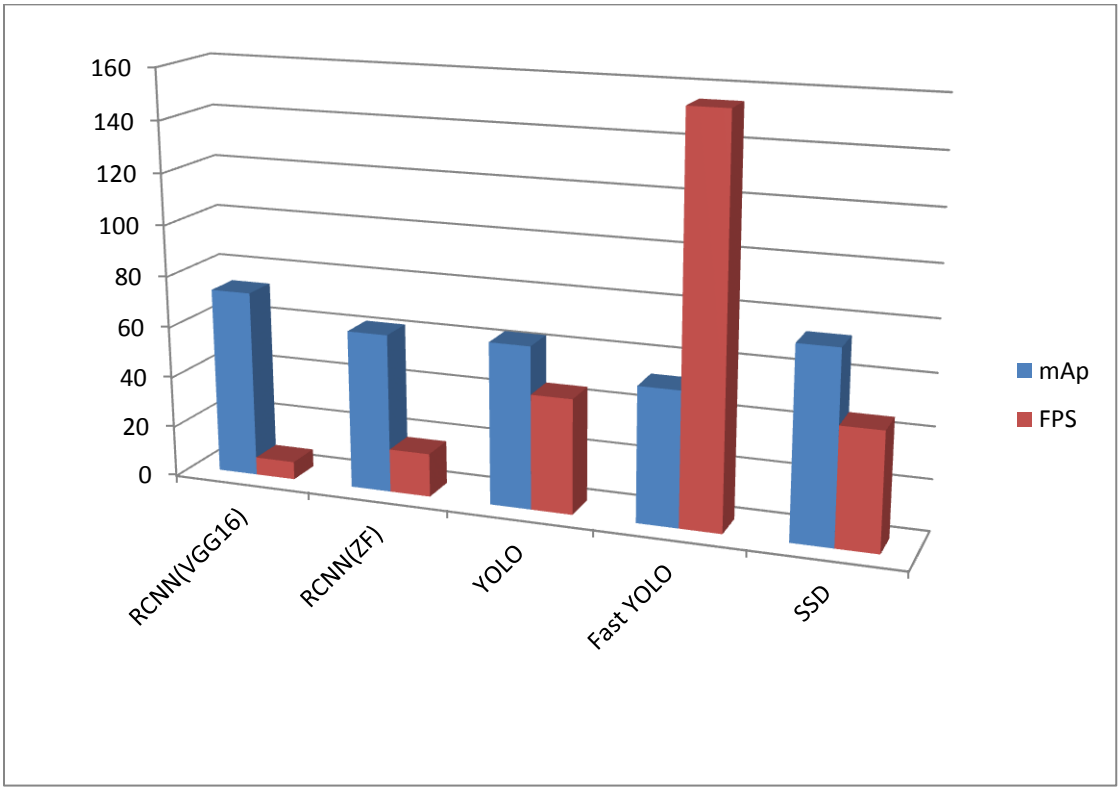


Fig.6.3.Comparison of different object detection models w.r.t to mAp and FPS

To obtain an optimum frame size for the system, series of experiments were performed by varying the resolution of the input image and the results are noted in Fig.6.4. The results show that the frame size of 256x256 performed the best. This frame size was chosen for carrying out further experimentations as it provided an acceptable level of accuracy in both faces and object recognition without exerting extensive overhead on the system.

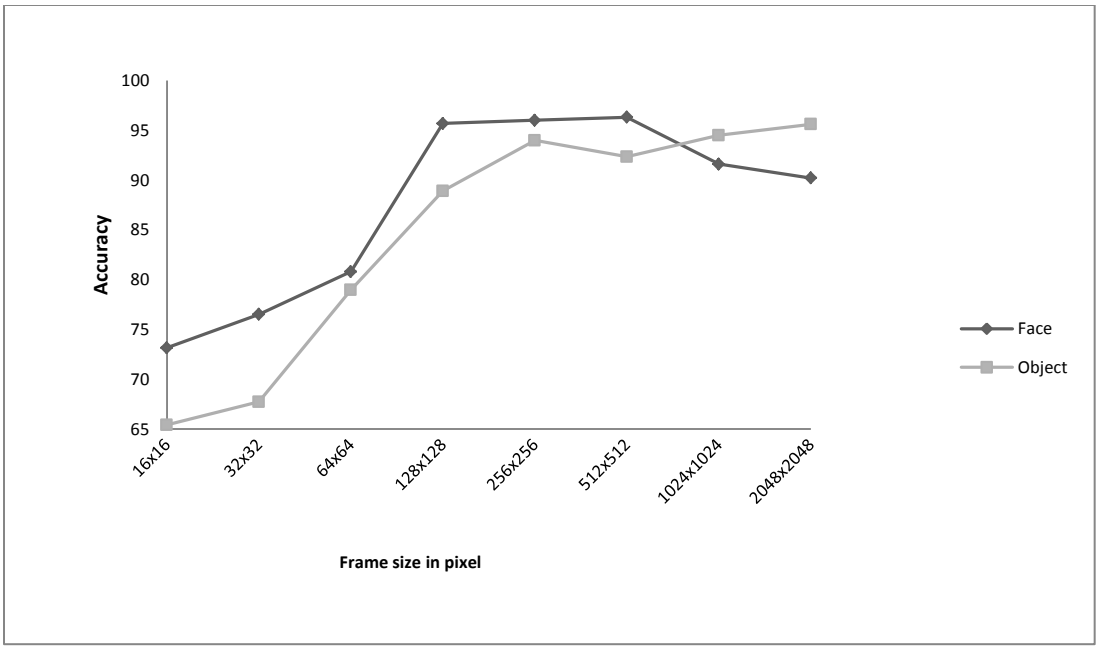
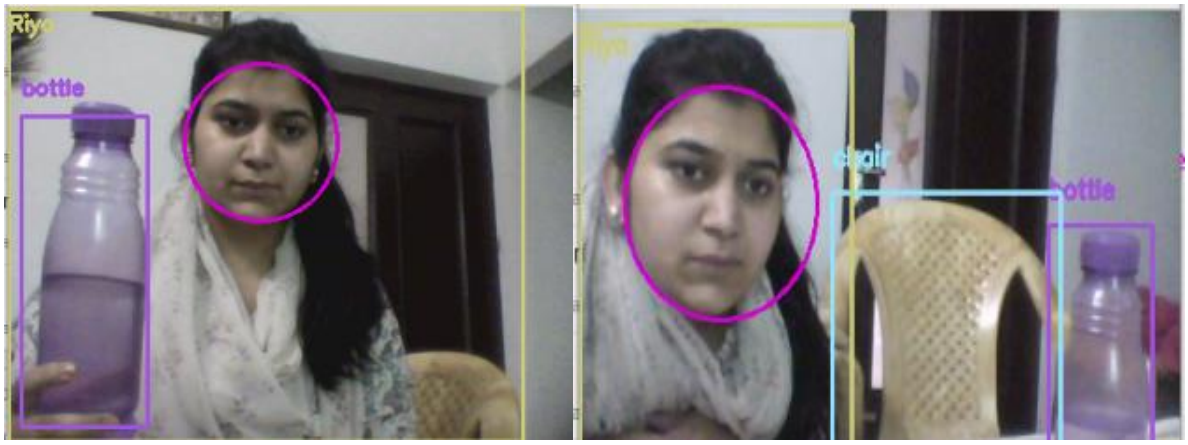


Fig.6.4. Accuracy v/s frame resolution of face and object recognition

The output of the system is shown in Fig.6.5. Here, multiple objects are identified using the chosen object detection algorithm. The corresponding labels and bounding boxes of these objects are also displayed to the user. If the system detects an object named as person then the face recognition model identifies the identity of that person using the corresponding face detection algorithm and displays his / her name.



(a) (b)

Fig.6.5.Frame showing multiple objects recognition (a) Person and Bottle (b) Person with Chair and Bottle

6.3 System Comparison and Evaluation

The result section of this approach comprises of reviews taken from multiple visually challenged people on one-week usage of the device. The device is tested on multiple visually impaired people in real time and feedbacks are recorded as shown in Table 6.3. A sample of 20 random users is selected for recording responses.

Table 6.3 shows that the device performs well with a high accuracy of 95.76%. Few users gave the feedback that the device adds additional weight to the white cane which affects its usage over a long span of time. Most of the users become quickly familiar with the device as it was fully automated and doesn't require any learning curve. The device performed exceeding well with 100% accuracy for few users as they were looking directly in the camera. Pose angle and intensity of light effected users 5,7,17 results giving less accuracy. Proposed device was also compared with other existing systems in Table 6.4.

Table 6.3.Results and feedback from the user

Users	**No. of Faces**	**No. of correctly identified faces**	**Accuracy**	**Is it found helpful for the user?**	**Feedback**
User 1	100	96	96%	Yes	No additional net charges is good
User 2	120	112	93.3%	Yes	Cheap and affordable
User 3	90	90	100%	Yes	Fast real time response
User 4	100	98	98%	Yes	Easy to use
User 5	150	135	90%	No	Adds additional weight to cane
User 6	100	93	93%	No	Charging Hassel
User 7	130	117	90%	Yes	Appreciated
User 8	115	112	97.4%	Yes	Really helpful
User 9	120	120	100%	Yes	Nice product
User 10	100	97	97%	Yes	Find easy to recognize objects of daily needs

User 11	95	93	97.9%	Yes	Portable
User 12	95	90	94.7%	Yes	Easy to carry around
User 13	100	99	99%	Yes	Appreciate the OCR for book reading
User 14	110	100	90.9%	Yes	Complete product
User 15	100	97	97%	No	No Comments
User 16	100	95	95%	Yes	Effective product
User 17	100	92	92%	Yes	Good sound interface
User 18	90	90	100%	Yes	Availability of regional language is helpful
User 19	95	91	95.8%	Yes	Nice
User 20	120	118	98.3%	Yes	All in one device

Table 6.4.Comparison of proposed application with the existing system

Author	**Internet Reliability**	**Object Recognition**	**Face Recognition**	**OCR**	**Face Recognition Accuracy**	**Object Detection Accuracy**	**OCR Accuracy**	**Indoor/ Outdoor**
Blasch, B.B et al.(1997) [5] and Kumar, K et al. (2014) [6]	No	Yes	No	No	-	50%	-	Both
Roshni et al. (2013) [7]	Yes	Yes	No	No	-	90%	-	Indoor
Nandhini et al. (2014) [8]	Yes	Yes	No	No	-	92%	-	Both
Simoes et al. (2012) [17]	Yes	Yes	No	No	-	93.5%	-	Indoor
Verma et al.	No	No	Yes	No	90%	-	-	-

(2015) [29]								
Kumar et al. (2014) [30]	No	No	Yes	No	85%	-	-	-
Deep learning based walking solution for visually impaired	No	Yes	Yes	Yes	95.76%	98%	92%	Both

The proposed device provided the mean average precision of 70% and accuracy of about 93.98% in object detection, the accuracy of 96% in face recognition and 92% in Object Character Recognition for both indoors and outdoors. A proposed system without using any high-end sensors provides a high recognition rate making it a portable, cheap, self-contained walking solution.

Chapter Summary

This chapter comprises of three subsections. Firstly, the evaluation metrics used for the analysis of each module is explained along with their mathematical formulation. The next subsection consists of experimentation results of face recognition and object detection model. In this a comparison between different face detection and object detection is shown in tabular form to analyze the best performing model for the system. Lastly, the system comparison and evaluation section shows the feedback gathered from various users and compared the proposed assistive device with the existing system.

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

This research presents a complete assistive solution for visually impaired people. The proposed assistive device consists of three major modules namely object recognition, face recognition and OCR. The object detection part of the approach helps in recognition of several objects like chair, table, TV etc. Face recognition module helps in authentication and recognition of the identity of a person in front of the user making him feel more secure.

The major problem faced by visually impaired is that they are constrained towards only a braille compatible book which in turn limits their knowledge acquisition. The OCR part of the framework empowers the user to grasp knowledge from any book of their interest. All the corresponding labels from tree modules are feed to text to speech module that provides audio feedback to the user from the ear piece. Entire system is integrated over raspberry pi board which is highly power efficient and portable. This complete system can be incorporated in a white cane or worn as a neck band acts as a third eye for the blind easing their daily lives.

7.2 Future Scope

There are certain limitations in the proposed system which can be resolved in future. The proposed assistive device is in its infancy hence it requires a lot of fabrication and efforts to make device feasible for the end user. Machine learning algorithms used in the approach can be further optimized to make it energy efficient. For object detection module two main improvements can be made, firstly to reduce the time of processing GPU's can be applied secondly the recognition module consists of only 21 classes which can be increased to include all relevant objects. Feedback acquired from multiple users can be used to make the system more user-friendly. It is also essential to make more experiments to further approve our method.

REFERENCES

1. World Health Organization, Visual impairment, and blindness: Fact sheet number 282, Aug. 2014. [Online]. Available: http://www.who. int/mediacentre/factsheets/fs282/en/, accessed on: Oct. 10, 2015.
2. Joseph, S. L., Xiao, J., Zhang, X., Chawda, B., Narang, K., Rajput, N., ... & Subramaniam, L. V. (2015). Being aware of the world: Toward using social media to support the blind with navigation. *IEEE transactions on human-machine systems*, *45*(3), 399-405.
3. Kef, S., Hox, J. J., & Habekothe, H. T. (2000). Social networks of visually impaired and blind adolescents. Structure and effect on well-being. *Social Networks*, *22*(1), 73-91.
4. Wang, C. W., Chan, C. L., Ho, A. H., & Xiong, Z. (2008). Social networks and health-related quality of life among Chinese older adults with vision impairment. *Journal of Aging and Health*, *20*(7), 804-823.
5. Wiener, W. R., Welsh, R. L., & Blasch, B. B. (2010). *Foundations of orientation and mobility* (Vol. 1). American Foundation for the Blind.
6. Kumar, K., Champaty, B., Uvanesh, K., Chachan, R., Pal, K., & Anis, A. (2014, July). Development of an ultrasonic cane as a navigation aid for the blind people. In *Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on*(pp. 475-479). IEEE.
7. Kumar, P. M., Gandhi, U., Varatharajan, R., Manogaran, G., Jidhesh, R., & Vadivel, T. (2017). Intelligent face recognition and navigation system using neural learning for smart security in Internet of Things. *Cluster Computing*, 1-12.
8. Nandhini, N., Vinothchakkaravarthy, G., Deepa Priya, G.: Talking assistance about location finding both indoor and outdoor for blind people. In: International Journal of innovative Research in Science, Engineering and Technology, vol. 3, pp. 9644–9651 (February 2014).
9. Dharani, P., Lipson, B., & Thomas, D. (2012). RFID Navigation system for the visually impaired. *Worcester Polytechnic Institute*, 48.
10. Koley, S., & Mishra, R. (2012). Voice operated outdoor navigation system for visually impaired persons. *International Journal of Engineering Trends and Technology*, *3*(2), 153-157.

11. Manogaran, G., & Lopez, D. (2017). Spatial cumulative sum algorithm with big data analytics for climate change detection. *Computers & Electrical Engineering*.
12. Manogaran, G., Thota, C., & Lopez, D. (2018). Human-computer interaction with big data analytics. In *HCI challenges and privacy preservation in big data security* (pp. 1-22). IGI Global.
13. Thota, C., Manogaran, G., Lopez, D., & Vijayakumar, V. (2018). Big data security framework for distributed cloud data centers. In *Cyber Security and Threats: Concepts, Methodologies, Tools, and Applications* (pp. 589-607). IGI Global.
14. Priyan, M. K., & Devi, G. U. (2017). Energy efficient node selection algorithm based on node performance index and random waypoint mobility model in internet of vehicles. *Cluster Computing*, 1-15.
15. Kumar, P. M., & Gandhi, U. D. (2017). A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases. *Computers & Electrical Engineering*.
16. Kumar, P. M., & Gandhi, U. D. (2017). Enhanced DTLS with CoAP-based authentication scheme for the internet of things in healthcare application. *The Journal of Supercomputing*, 1-21.
17. Simões, W. C. S. S., & de Lucena, V. F. (2016, January). Blind user wearable audio assistance for indoor navigation based on visual markers and ultrasonic obstacle detection. In *Consumer Electronics (ICCE), 2016 IEEE International Conference on* (pp. 60-63). IEEE.
18. Lakde, C.K., Prasad, P.S.: Navigation system for visually impaired people. In: International Conference on Computation of power, energy, Information, and Communication (2015)
19. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).
20. Sun, Y., Wang, X., & Tang, X. (2013, June). Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on* (pp. 3476-3483). IEEE.
21. Zhang, C., & Zhang, Z. (2010). A survey of recent advances in face detection.
22. Jafri, R., & Arabnia, H. R. (2009). A survey of face recognition techniques. *Jips*, *5*(2), 41-68.

23. Timo, A. (2004). Face recognition with local binary patterns. In *Euro. Conf. on Computer Vision*.
24. Chen, X., Flynn, P. J., & Bowyer, K. W. (2005). IR and visible light face recognition. *Computer Vision and Image Understanding*, *99*(3), 332-358.
25. Lei, Y., Bennamoun, M., Hayat, M., & Guo, Y. (2014). An efficient 3D face recognition approach using local geometrical signatures. *Pattern Recognition*, *47*(2), 509-524.
26. Lopez, D., & Manogaran, G. (2017). Modelling the H1N1 influenza using mathematical and neural network approaches. *Biomedical Research*.
27. Manogaran, G., Thota, C., Lopez, D., & Sundarasekar, R. (2017). Big data security intelligence for healthcare industry 4.0. In *Cybersecurity for Industry 4.0* (pp. 103-126). Springer, Cham.
28. Manogaran, G., Lopez, D., Thota, C., Abbas, K. M., Pyne, S., & Sundarasekar, R. (2017). Big data analytics in healthcare internet of things. In *Innovative healthcare systems for the 21st century* (pp. 263-284). Springer, Cham.
29. Verma, R. N., Jain, K., & Rizvi, M. A. (2015, September). Efficient face recognition method using RBF kernel and genetic algorithm. In *Computer, Communication and Control (IC4), 2015 International Conference on* (pp. 1-5). IEEE.
30. Kumar, A. L., & Ganesan, R. (2014, March). Improved navigation for visually challenged with high authentication using a modified sift algorithm. In *Green Computing Communication and Electrical Engineering (ICGCCEE), 2014 International Conference on* (pp. 1-5). IEEE.
31. Raghavendra, R., Rao, A., & Kumar, G. H. (2010). Multimodal person verification system using face and speech. *Procedia Computer Science*, *2*, 181-187.
32. Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433-459.
33. R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of eugenics 7 (2) (1936) 179 – 188.
34. D. Gavrila, S. Munder, Multi-cue pedestrian detection and tracking from a moving vehicle, International Journal of Computer Vision 73 (1) (2007) 855 41–59.
35. Goel V, Mishra A, Alahari K, Jawahar C V. Whole is greater than sum of parts: Recognizing scene text words. In: Proceedings of IEEE International Conference on Document Analysis and Recognition. 2013, 398–402

36. Jaderberg M, Simonyan K, Vedaldi A, Zisserman A. Reading text in the wild with convolutional neural networks. International Journal of Computer Vision, 2016, 116(1): 1–20
37. Wang T, Wu D J, Coates A, Ng A Y. End-to-end text recognition with convolutional neural networks. In: Proceedings of IEEE International Conference on Pattern Recognition. 2012, 3304–3308
38. Mishra A, Alahari K, Jawahar C V. Top-down and bottom-up cues for scene text recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2012, 2687–2694.
39. He P, Huang W, Qiao Y, Loy C C, Tang X. Reading scene text in deep convolutional sequences. In: Proceedings of AAAI Conference on Artificial Intelligence. 2016
40. Shi B G, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. 2015, arXiv preprint arXiv:1507.05717.
41. http://cnyah.com/2018/01/03/computer-vision-project-summary/
42. http://cv-tricks.com/object-detection/faster-r-cnn-yolo-ssd/
43. Ueki, K., Kobayashi, T.: Multi-layer feature extractions for image classification—Knowledge from deep CNNs. In: 2015 International Conference on Systems, Signals, Image Processing (November 2015).
44. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS (2015).
45. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multi-box detector. In *European conference on computer vision* (pp. 21-37). Springer, Cham.
46. https://docs.opencv.org/3.4.1/d7/d8b/tutorial_py_face_detection.html

LIST OF PUBLICATIONS

1. K. Kalra, R. Goyal, S. Kaur, P. Bhatia, "PDD Algorithm for Balancing Medical Data" in 2nd International Conference on Advances in Computing and Data Sciences (ICACDS 2018)

   [Accepted]

2. R. Goyal, K. Kalra, P. Bhatia, S. Kaur, "Intelligent Face Recognition System For Visually Impaired" in 2nd International Conference on Advances in Computing and Data Sciences (ICACDS 2018).

   [Accepted]

3. R. Goyal, K. Kalra, P. Bhatia, S. Kaur, "Automated real time hand gesture recognition interface" in Journal of multimode user interface (Springer).

   [With Editor]

4. K. Kalra, R. Goyal, S. Kaur, P. Bhatia, "DESIRE: Deep Learning Enabled System Integration for Retinopathy Evaluation" in International Journal of pattern recognition and artificial intelligence (World Scientific).

   [With Editor]

5. R. Goyal, K. Kalra, P. Bhatia, S. Kaur, "UMEED: Deep Learning based complete walking and reading solution for visually impaired people" in Universal Access in the information Society (Springer).

   [Communicated]

VIDEO LINK

Riya Goyal, “UMEED: Walking Assistant for Blind”, https://youtu.be/ceWm9bt35vI