Indian Journal of Science and Technology, Vol 9(30), DOI: 10.17485/ijst/2016/v9i30/98716, August 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Camera based Text to Speech Conversion, Obstacle and Currency Detection for Blind Persons

D. B. K. Kamesh*, S. Nazma, J. K. R. Sastry and S. Venkateswarlu

Department of Electronics and Computer Engineering, KL University, Vaddeswaram - 522 502, Guntur District, Andhra Pradesh, India; kameshdbk@kluniversity.in, shaiknazu321@gmail.com, drsastry@kluniversity.in, somu23@kluniversity.in

Abstract

Background/Objectives: The main object of this paper is to present an innovated system that can help the blind for handling currency. **Methods/Statistical Analysis**: Many image processing techniques have been used to scan the currency, remove the noise, mark the region of interest and convert the image into text and then to sound which can be heard by the blind. The entire system is implemented by using Raspberry Pi Micro controller based system. In the proto type model an IPR sensor is used instead of camera for sensing the object. **Findings:** In this paper a novel method has been presented using which one can recognize the object, mark the interesting region within the object, scan the text and convert the scanned text into binary characters through optical recognition. A second method has been presented using which the noise present in the scanned image is eliminated before characters are recognized. A third method that can be used to convert the recognised characters into e-speech through pattern matching has also be presented. **Applications**: An embedded system has been developed based on ARM technology which helps the blind persons to read the currency notes. All the methods presented in this paper have been implemented within an embedded application. The embedded board has been tested with different currency notes and the speech in English has been generated that identify the value of the currency. Further work can be done to generate the speech in different other both National and International Languages.

Keywords: Camera based Detection for Blind Persons, Currency Detection, Raspberry Pi Board, Text to Speech Conversion

1. Introduction

A camera based reader helps blind persons to read labels on the products and other handheld devices in their day by day lives[1]. To differentiate the object from heavy backgrounds and other surroundings, an effective motion based method is used to define Region of Interest (ROI) in the camera view. In the obtained ROI text, localization and recognition are done[2]. The printed context or scanned image is converted into computer recognition format by using Optical Character Recognition (OCR) so that it can increase the speed of operation[3]. In model identification, all the characters are bounded and detached and then the end character image is directed to a pre-processor for removing the noise. All the characters are compared with a database of identified characters which are assembled together to form initial text pattern. The output is then given to the e-speak engine to convert text to speech and this output is given to blind users through earphones. The obstacle in the process is identified by using PIR sensor.

285 million people are evaluated to be visually impaired worldwide, 39 million are blind and 246 million have low vision[4]. This paper mainly offers a low cost system to help the blind persons. The latest progress in digital cameras, portable computers helps in designing the camera based products that combines the computer vision technology and the optical character recognition system. Software's such as video magnifiers, screen readers and optical aids are available to help the blind people and those with vision loss to use a computer. There are only few devices that can offer good access to blind users to read printed text in outside world. In today's world, number of challenges have been faced by the blind people because printed text is everywhere in the form of receipts,

*Author for correspondence

bank instructions over the medicines etc. When blind people are aided to read printed text and product labels, it will increase their independent living as well as foster social and economic self-efficiency.

Existing systems such as movable bar code readers constructed to name various products in an extended data base can facilitate the blind users to route the information about these products over speech[5,6]. On the other hand a gigantic drawback in that is it is extremely tough for blind users to locate the place of the barcode and exactly point the barcode reader at the barcode. Another reading assistive system such as pen scanner is developed in these analogous circumstances[7]. Other major problems for blind people are obstacle detection and identifying different denominations of the currencies. There are almost 50 currencies all over the world and each of them looking totally different. In the prototype system, a camera is used for reading the text and the currency. Text is converted into different languages such as English, Tamil, Telugu, Urdu[8,9]. The entire application was run on Raspberry Pi board.

2. Text to Speech Conversion

2.1 Software and Hardware Specifications

To implement text to speech conversion, a platform has been selected which include Operating system (Linux), Language (python), Library (Open CV (Linux-library)), CPU: 700 MHZ low power, ARM1176JZ-F applications processor, 512 MB SDRAM, micro SD of 8 GB, multi-channel HD audio over HDMI, USB 2.0 and supported 640×350 to 1920×1200 including 1080 p.

2.2 Raspberry Pi

The Raspberry Pi board used for development and experimenting the methods presented in this paper is shown in the Figure 1. It is a mini computer which is in credit-card size that plugs into a TV or monitor and uses a mouse and keyboard. It allows people of all age groups to use a computer and help them in learning programming languages like python and scratch.

2.3 Block Diagram

The text to speech framework consists of three essential mechanisms Image capturing, Data managing and speech output. The image capturing element gathers the images which contains text in the form of video or image. In our prototype system images are captured by using a WEB CAM. Using the camera, image of the object from the cluttered backgrounds or other surroundings are extracted. Text localization algorithm is used for acquiring the text from images; an Ada boost learning model is applied to localize the text in the image. Text recognition is implemented to convert the image based text to readable codes. In the prototype developed a laptop is used for data processing. The audio output is delivered to the blind users through e-speak engine and the audio is presented in English language. The block diagram that shows various parts of the system for implementing the proposed methods has been shown in the Figure 2.

Figure 1. Raspberry Pi board.

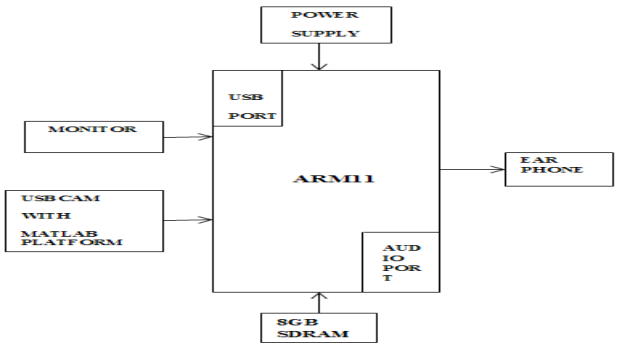


Figure 2. Proposed block diagram.

2.4 Character Recognition System

Images for Character Recognition (CR) system have been acquired by capturing the photograph of a document or product or by scanning the handwritten text or by immediately writing into a computer system. The noise available in the scanned object is removed at pre-processing stage at the time of text creation, appropriate filters such as min-max filter, Gaussian filter etc. are used to eliminate the noise. "Binarization" method changes black and white

or coloured image to binary image which is in the form of black or white. Generally grey image is the combination 0s and 1s and the binary image is the combination of 0s or 1s. The positive values above the threshold level are taken as 1s and the background with negative values are taken as 0s. Almost image processing method output is in the form of binary image. On thresholding, a grey level image with pixel values ranging between 0 and 255 represent the binary image[10]. While scanning the text, it may or may not be exactly horizontally placed, by using the slant angle correction it is exactly aligned. If the input image is too large then it is resized to diminish the dimensions to upgrade the speediness of processing. During segmentation all lines are separated by applying row histogram and by using column histogram, words are extracted from each row and last characters are extracted from words. Feature extraction is an essential part of any pattern recognition application and features of individual characters are extracted, wavelet based multiresolution technique for feature extraction has been used. Classifiers match the input characteristics with gathered sample and identify perfect equivalent input. Post processing improves the accuracy of recognition and therefore needs to be used sometimes. The step by step process used for recognizing the characters has been shown in the Figure 3.

Image Problem

Figure 3. Character recognition systems.

3. Object Detection

In prototype system presented, object is detected by using PIR sensor; it is connected to the Raspberry Pi board. When any body's hand is placed in front of that PIR sensor, e-speak engine alerts the blind person that a person is ahead[11,12].

3.1 Passive Infrared (PIR) Sensor

It is a pyro electric machine that identifies movement by computing variations in the infrared intensities by nearby objects. This motion can be sensed by examining the strong signal on a single input/output pin. It has the ability to reliably separate movable bodies from other objects as well as from stationary bodies. It is very small in size that makes it make it easy to conceal in security applications also. It has 3.3 and 5 v operation with <100 μA current consumption.

4. Currency Detection

The system presented in this paper is mainly based on image processing, filtering, edge detection, segmentation. To make the system more comprehensive, it is necessary to create a database for storing the characteristics of currency. In the system presented, Indian currency is used as example. The system is programmed using MATLAB. Several steps of processing have been implemented which include: 1. Reading the image obtained from the scanner and it is in JPEG (Joint Photographic Experts Group) format. 2. Pre-processing the image, removing the noise included and smoothening. 3. Edge detection, segmentation, pattern matching. Pre-processing of the image is done at low level of abstraction, so histogram equalization is used for increasing the image clarity by modifying the brightness and contrast. By comparing the currency with database, fake currency is also determined by using the PCA algorithm. This algorithm mainly used to extract the non-linear characteristic between different variables and separate the key features of data[13].

The hardware connections that have been established using the experimental model for achieving the currency detection and for converting text to speech is shown in the Figure 4.

The hardware design includes Memory card slot of 8 GB, Power supply, Mouse and keyboard connection, Ear phone, HDMI VGA cable connected to monitor and

Raspberry pi board, USB-2 serial port, MAX 3232 and Web CAM. The Input Image, the image taken from camera and the character recognition obtained as output from the Experimental model is shown in Figure 5, Figure 6 and Figure 7 respectively.

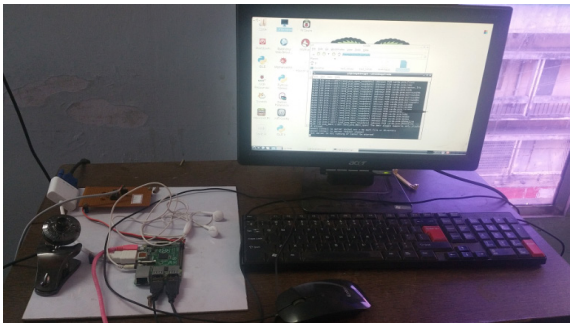


Figure 4. Hardware connections for text to speech.

Figure 5. Input images.

Figure 6. Image taken from camera.

The entire project runs on Linux library and MATLAB. In the above hardware connection Vcc is connected to pin 2 and ground to pin 6, transmitter connected to pin 10 and receiver to pin 8 on Raspberry Pi.

The input image taken is in grey form and then converted to binary so that character separation is done and each and every character is compared with the data contained in the database to find the authentication of the data. The data representing the image is fed to an e-speak engine to convert text to speech which is delivered to the blind user through ear phones. The hardware connections made for detecting the obstacles are shown in the Figure 8.

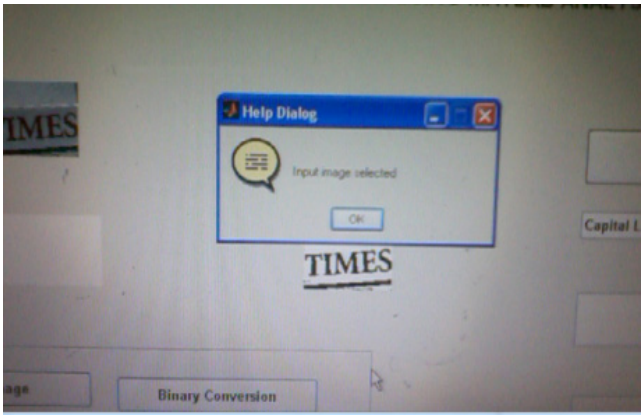


Figure 7. Character recognition.

Figure 8. Hardware connections for obstacle detection.

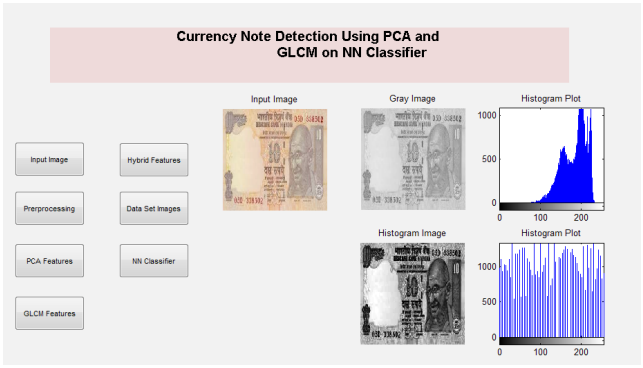


Figure 9. MATLAB executions for currency detection.

For obstacle detection PIR sensor is connected to the Raspberry Pi board when the obstacle is placed in front of

this sensor it gives information to the blind person that "A Person Is Ahead". The range of this PIR sensor is 7 meters. The hardware connection for currency and text to speech is same. The input image for currency detection is shown in Figure 9.

5. Conclusions

The proposed system helps the vision impaired people in their daily activities. This model helps to read the text and converts it in to audio form. It also helps to find out the obstacle by using PIR sensor and to find out the currency through use of a database. The PIR sensor alerts the blind user about the obstacle but cannot name the object. In future it can be further extended by using the latest obstacle detecting technologies and also helps in identifying the currency in real time.

6. References

1. Yi C, Tian Y, Arditi A. Portable camera based assistive text and product label reading from hand-held objects for blind persons. IEEE/ASME Transaction on Mechatronics. 2014 Jun; 19(3):808–17.
2. Raj Kumar N, Anand MG, Barathiraja N. Portable camera based product label reading for blind people. IJETT. 2014 Apr; 10(11):521–4.
3. Mohammad F, Anarase J, Shingote M, Ghanwat P. Optical character recognition implementation using pattern matching. International Journal of Computer Science and Information Technologies. 2014; 5(2):2088–90.
4. World Health Organization. 10 facts about blindness and visual impairment. 2015. Available from: http://www.who.int/features/factfiles/blindness/blindness_facts/en/
5. Reading ID barcodes with mobile phones using deformable templates. 2015. Available from: http://a4academics.com/be-seminar-topics/17-be-it-cse-computer-science-seminar-topics/426-reading-1d-barcodes-with-mobile-phones-using-deformable-templates
6. Introduction to Braille language. 2016. Available from: http://www.acharya.gen.in:8080/disabilities/br_intro.php
7. Pen with scanner. 2016. Availale from: http://www.prv.se/en/patents/why-apply-for-a-patent/examples-of-patents/pen-with-scanner
8. Rajendran V, Bharadwaja Kumar G. Text for developing unrestricted tamil text to speech synthesis system. Indian Journal of Science and Technology. 2015 Nov; 8(29):1–10.
9. Jyothi J, Manjusha K, Anandkumar M, Soman KP. Innovative feature set for machine learning based Telugu character recognition. Indian Journal of Science and Technology. 2015 Sep; 8(24):1–7.
10. Binary Images. 2016. Available from: http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/OWENS/LECT2/node3.html
11. Sensing Motion with Passive Infrared Sensors. 2016. Available from: http://www.digikey.com/en/articles/techzone/2012/jun/sensing-motion-with-passive-infrared-pir-sensors
12. Raj KM, Raj MS. Person identification for visually impaired using lab VIEW. Indian Journal of Science and Technologies. 2015 Nov; 8(31):1–4.
13. Vashishtha V, Md S. A paper currency recognition system using image processing to improve the reliability with PCA method. International Journal of Engineering Sciences and Research Technologies. 2015 Jun; 4(6):172–5.

Artificial Eye for the Blind

Abhinav Benagi , Dhanyatha Narayan , Charith Rage , A Sushmitha
Computer Science Enginnering , M S Ramaiah Institute Of Technology, India

Abstract

Visual Impaired humans cannot perceive their environment and navigate like normal Humans do , which results in reduced mobility.They also fear the society and are often treated as incable submissive section of society. In the light of making a blind person overcome his fear and step out to the society independently and confidently, we have added a few features to the blind stick enabling the blind to be able to detect gender,age and basic actions done by a person through which he can react to his opposite party in an appropriate manner.

This project is useful when it comes to blind people being able to detect facial expressions in order to respond to environmental situations precisely.It also helps them detect the age and gender of the person to respond accordingly. It also helps them interpret text from images without relying on the third party for help.For security purposes, it helps in detecting the actions carried out by the person in the surroundings and taking appropriate actions against the situation. If this project is implemented appropriately, it can help blind people in carrying out a wide range of everyday activities effortlessly which adds on a lot of scope for the juvenile generation.

For facial expression detection, AI uses a lot of non-verbal functionalities such as facial expressions, body language and gestures. Whereas, to determine the age and gender it uses a broad dataset which includes a significant amount of pictures of men and women contributing to the dataset which in turn helps in accurately determining the age and gender of the person.Text to speech conversion is done by storing many different font and text images in patterns. It uses matching algorithms to distinguish between text images, character by character.In conclusion, the blind stick accommodates multiple features each of which can be accessed with separate functionalities such as, one functionality for action and object detection , one for image to speech conversion , one for facial recognition.

Table of content

Chapter 1 - Introduction

Globally, Around 2.2 Billion people don't have the capability to see and 90% of them coming from low-income countries tell us that, it is a need of the hour for an easily accessible, economically viable and ethically appropriate equipment for the specially abled. Unknown environments are especially challenging as it is difficult for a person to detect the objects in the surroundings or read any sign boards, understand the behavior of the human or navigate themselves to the destination.

Blindness is one of the most misunderstood types of disability ,as many people believe that a blind person cannot do their work or live normally. Millions of them are in India and are facing troubles in their daily lives as we don't have the proper equipment for them. Adding on, common able-people are often judgemental about the troubles faced by the visually impaired which makes them hesitant to face the dominant world.The objective of this project is to ensure that the visually impaired will now emerge to be more confident and feel more empowered as compared to other sections of the population. With the implementation of this project, the blind can now be less dependent on their current environment and people.They can now get first hand access to technology and make best use of it.

There are several devices and aids for the blind that are already available But, all of them focus on a specific functionality. Our project comes with a solution that can accommodate provisions for multiple functionalities during the same computational period making it stand out amongst the existing devices. Braille is one of the places the specially abled can comfortably read in different languages too. But, it too is not available everywhere across the globe thus pushing them towards a means for reading and recognizing text. This section of the society, being more prone to fatalities on the road accidents need a feasible device through which they can stay afar from dangers.

Hence, in this project we have included features like object detection,image to text conversion and then to speech,gender and age detection and ultimately action detection. Furthermore it can also be converted to speech so that the blind person can hear using his earphone. When a blind person needs to know directions by reading the sign boards or when he has to read a non-braille book he can just click the picture of it using the feature of image to text conversion where, the captured image can be analyzed and the English text which is present can be recognised and by using text to speech conversion feature the text is read out and can be understood by the blind. A sensor connected to the blind stick beeps when the object is in the range of specified distance.

Action detection is the task of identifying when a person in an image or video performs atomic actions such as walking ,bending,clapping,falling etc. This helps a blind person to evaluate the situation and react accordingly.It is vital for any species to know what is going on and how the world revolves around them. When you are visually impaired, chances of your other senses becoming extremely sharp are really high. Most blind people would automatically develop sharp hearing senses which is a good sign of empowerment but it's not necessary that only improved senses can help a blind get through any difficult situation. Although it's not impossible to deal with situations without action detection it is sometimes really helpful if the blind can detect what actions are being done by what people around him. For instance let us consider there's a man putting up an act in an open space environment, it would be helpful for the blind to know what's happening around to be aware of any dangerous acts occurring around him.

The methodology used in object and action detection is, capturing a real time snapshot of the present situation and using tensor flow modules and libraries to detect what activity the respective person is doing and what object he/she is holding.This can be done using video detection also instead of capturing snapshots by feeding the program with a pre recorded video. Alternatively, real time video detection was not chosen since the microcomputer cannot handle high GPU power. The mentioned features are implemented via raspberry pi microcontroller.

Object recognition is a classical problem in computer vision, determining if the image contains a specific object.so we chose this problem upon color recognition, shape recognition etc.This feature identifies objects present in real time more accurately as the dataset contains images in many different contexts .It can perform better at identifying objects in various environment.Object detection bounds boxes around these detected objects which allows us to identify the coordinates of each object exactly .

Optical character recognition(OCR) is an optical scanner, also a specialized software ,used to extract text from images . Google Text to Speech(gTTS) library in python is used,which converts text into audio and allows us to save this audio as a mp3 file. The speech can be delivered at various speeds and can be customizable based on the user's grasping level .

Age and gender detection is implemented using deep learning concept to accurately identify age and gender of a person from a single image of a face.It is very difficult to accurately guess an exact age from a single image because of

factors like makeup,lightning,obstruction an facial expressions so we make this a classification problem instead of regression. For all of us it is important that we are addressed appropriately, with respect to age and gender. A lady wouldn't want to be addressed as male and the older section of the society would not like to be disrespected. Therefore it is important for the visually impaired to be able to identify the age and gender of an individual so that he/she can fit into the society without much trouble.

Our Blind stick project implements multiple functionalities different from a normal blind stick which only beeps on the detection of an object in front of it, which is exactly the cause for why we think this project has a wide scope in future. Only when we walk in their shoes will we know the casualties and difficulties faced by them. In this project we have tried our level best in understanding the problems faced by the blind and tried implementing a solution for it.

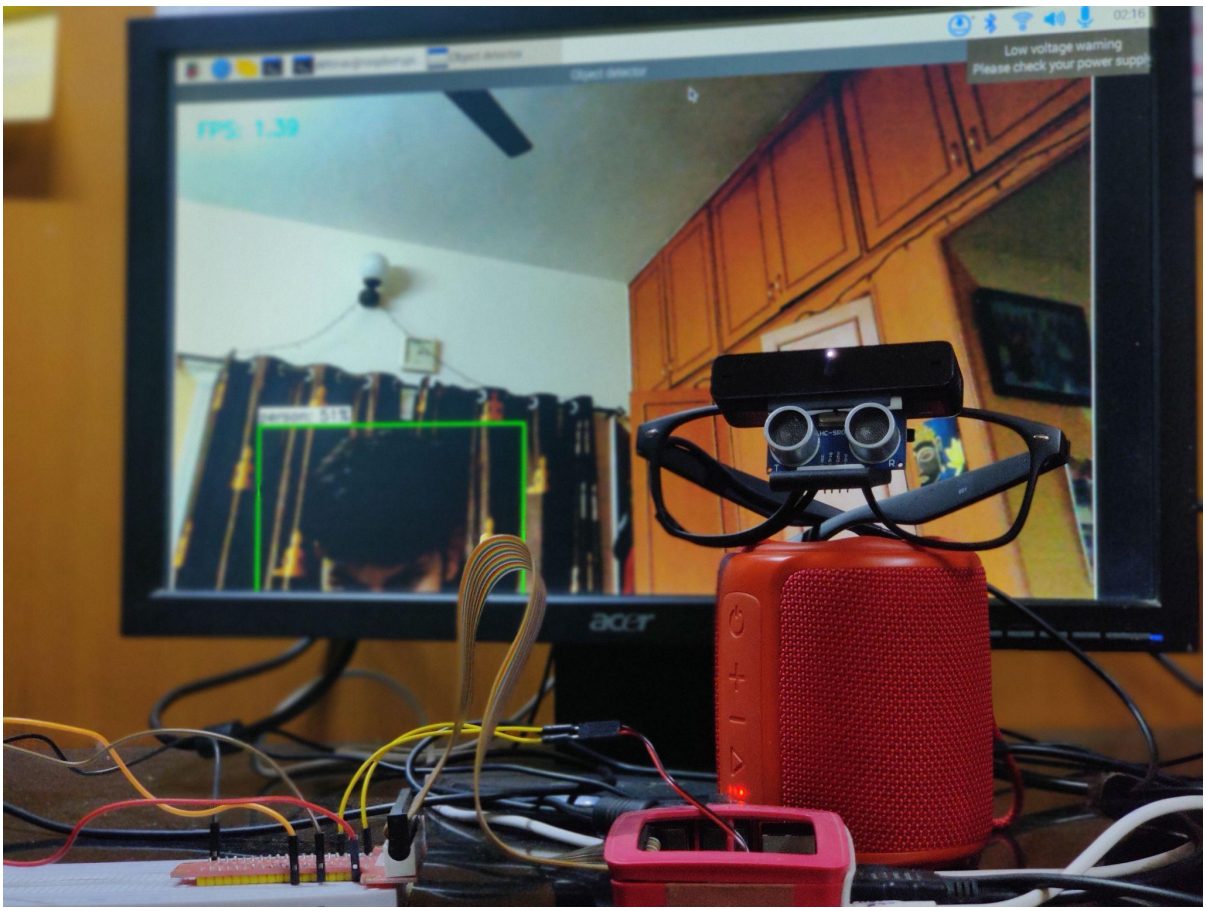


Fig 1: Artificial Eye

Chapter 2 - Literature Review

There are many advantages of the advanced driving assistance system(ADAS) for a better driving experience such as this system operates on radars, LiDARs for object detection and therefore providing our model high speed, low cost and lower power consumption. The proposed system used in this paper for object detection using YOLOv5 which helps us reduce the occlusion issue. We also study the services and benefits of darknet CNN and its 53 convolutional layers which when stacked onto the original architecture summing up to 106 layers. Darknet also acts as a backbone to the YOLOv5 algorithm as proposed in [1].

The recent developments of object detection using deep learning by building convolutional neural networks(CNN) was observed. We also learn the evaluation of tensor flow object detection framework for robust detection of traffic light [2]. It also tells us that faster RCNN delivers 97.015%, which outperformed a single shot multibox detector(SSD) by 38.806% for a model trained with a set of images.A case study about traffic lights in Malaysia and the images that were collected and stored as a dataset was also demonstrated. Detection of eyes and the sobriety of the person is done to determine whether the person is under the influence of alcohol. Safety in self-driving cars is improved using the model [2].

We understand that CNNs are one of the best algorithms to process the image content and have shown immense improvements in the image content segmentation, classification, detection and retrieval related tasks from [3]. CNN also has a feature to capitalize on spatial or temporal correlation in data. The nonlinearity generated in the different patterns of activation of different responses therefore gives a different semantic understanding for images. The ability to learn good representation from raw pixels without exhaustive processing, hierarchical learning, automatic feature extraction, multi-tasking and weight sharing makes it stand out amongst other algorithms. Backpropagation algorithm helps in the learning of the model by manipulating the change in the weights according to its respective target. Thus, we understand that deep architectures have an edge over shallow architectures due to its human level performance.

YOLO (you only look once) is a Powerful algorithm used for image processing through which objects can be detected with great precision and accuracy. A single convolutional network parallelly predicts bounding boxes and class probabilities for those boxes. The algorithm trains full images and optimizes our overall performance. This model has vast benefits over traditional methods of object detection. It could be compared with other powerful algorithms such as tensor

flow with respect to speed and accuracy. Its speed is extremely fast since detection is framed as linear regression and a simple pipeline is laid out [4]. Base network runs at 45 frames per second with no batch processing on a Titan X GPU and the fast version runs at more than 150 fps. This tells us that we can process streaming video in real-time with less than 25 milliseconds of latency[4]. We increase the precision by two times.

There are multiple sensing modalities like RGB cameras, depth cameras, inertial sensors etc through which actions of humans are determined. Fusion of the decisions obtained from two different modalities make the model more robust. Actions sometimes can be categorized into interest and non-interest actions from a very vast action stream is a primary necessity that has to be satisfied in an action detection model. There are several practical examples in the modern day world where actions dictate many tasks in our life. Gesture detection features are being incorporated into mobiles, TV's and any other automation where there is a need for human intervention, automations like these play a critical role in [5]. CNN(convolutional neural network) and Long short term networks(LSTM) based fusion systems are used to detect actions of interest from the stream.

Tesseract OCR Engine makes use of Long Short Term Memory (LSTM), a part of RCNN's[6]. It is suitable in recognizing larger portions of text data rather than single characters. Errors occurred are reduced during character recognition. It is difficult for visually impaired people to read textual information. The Blind have to make use of Braille to read. Instead it would be an easier task if they could simply listen to the audio version of the text. This application is a viable alternative to convert textual data to audio format. Google Text-To-Speech API facilitates this need.[6]

It is vital for the visually impaired to identify the activities performed by the people around them. This can be done through various convolutional models and tensorflow is one among them.The research [7] builds a human action recognition system based on a single image or a video captured. The TensorFlow Deep Learning models are developed using human keypoints generated by OpenPose. Four classifiers are considered: Neural Network, Random Forest, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) Classifiers.. The models' input layer is 50 points from x and y coordinates of 25 keypoints from OpenPose, and the output layer is the numerical representation of various human action labels which like hand waves, planks, running, sitting, hiding etc [7].

It is necessary for a person with visual aid to identify the gender and age in order to address the opposite party respectfully. Gender and age play a significant role in interpersonal interactions among people who live in communities. Image enhancement used in [8], is the process of improving an image so that the resultant image is of higher quality and can be used by other applications. The image is divided into a limited number of objects in order to solve the problem, this is called Segmentation. Due to the accuracy of its classification technique, deep learning techniques are a variety of tasks such as classification, feature extraction, object recognition, and so on. It helps in gender and age prediction.In the CNN model used in [8], first the face is extracted from a webcam image before proceeding with the implementation. The OpenCV library in Python is used to accomplish this. Haar feature-based cascade classifiers canbe used for detecting objects

Description, Feature extraction and gender classification is done sequentially in [9]. The input image is taken, then the central line location and various organs like eyes, noses, mouth are found. Feature extraction is done after locating all the components mentioned above with the help of PCA(principal component analysis). PCA helps us to predict, remove redundancies and compress the compiled data. It also tends to find a M dimensional subspace from a face image that is represented as a two dimensional N by N array of intensity values.The age of the person is determined with the help of a face space which is found out by computing the euclidean distance of feature points in two faces. Representation of data in columns of matrices and computing the covariance matrix help us to find the eigen vector and therefore allows us to determine the gender with accuracy.NNC(nearest neighbor classification) and KNN(kth nearest neighbor) classifiers are used to correctly classify the faces and determine the gender of the person.

Optical Character Recognition (OCR) is a branch of AI which is used to detect and extract characters from scanned documents or images and convert them to editable form. We then convert this extracted text to speech making it understandable to the visually impaired. Traditional methods of OCR used CNN's but they are complex and usually suitable for single characters. These methods also had a higher number of errors and less precision.[10]

Chapter 3 – Implementation

The main backbone of our Artificial Eye model is the Raspberry pi3 which is connected to the webcam ,ultrasonic proximity sensor, speaker and we also run all our software models i.e object detection, Optical Character recognition, google text to speech conversion and the Mycroft voice assistance model.

At first the ultrasonic proximity sensor will be measuring the distance between itself and any obstacle in front of it .When the Proximity sensor detects any obstacle in front within its specified range, the blind person will hear an audio prompt about an obstacle in his way at a certain distance. At this time the Webcam will capture an image in front of it and the Object detection model and the Optical Character Recognition model will begin to run on the Raspberry pi. The image captured is first sent through the Tesseract OCR module to detect any texts in the image and then through the Object detection model to detect the objects in front of the blind person. The text and the object detected are conveyed to the blind person by converting the texts to speech by using the gTTS module.

Along with the above mentioned process going on there will be an active MYCROFT voice assistant model which can be used to interact with the blind person. The blind person can ask about the weather , daily news , any information on the internet ,etc.

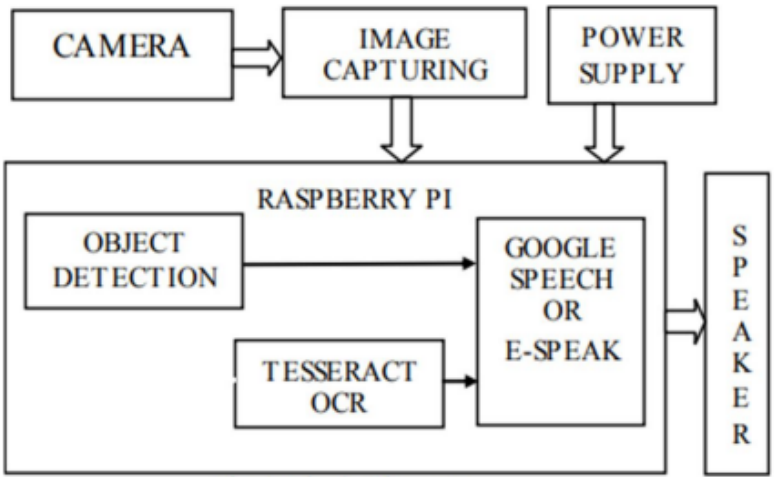


Fig 2: general working diagram

Proximity Sensor

We are using the Ultrasonic SR04 sensor as our proximity sensor.Ultrasonic sensors calculate's the obstales distance by emitting ultrasonic sound waves and converting those waves into electrical signals. The ultrasonic sensors have a range of up to 40-300cm with a great response time of up to 50 milliseconds to 200 milliseconds. This Proximity sensor's Vcc,gnd,Trig,Echo pins are connected to the Raspberry pi 3 b+ GPIO pins . We have used a python file in the raspberry pi for the working of this ultrasonic sensor.

The ultrasonic sensor keeps transmitting signals continuously , when an obstacle comes in front of it the receiver receives the signals and the obstacle is detected. Once the obstacle is detected the python script in the raspberry pi converts the distance into an audio format telling the blind person the distance between himself and the obstacle. We have used the Google Text to Speech api to convert the string to an audio output.

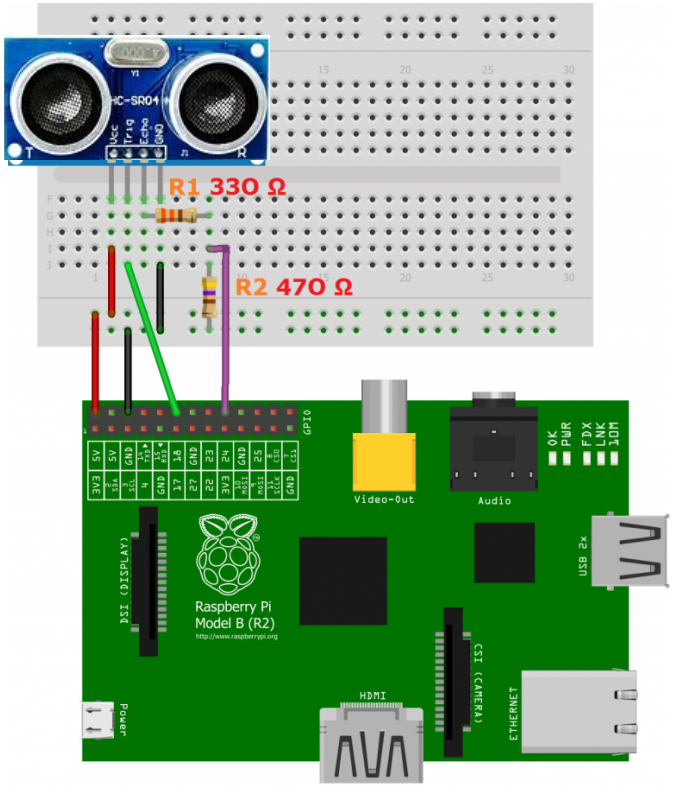


Fig 3. SR04 ultrasonic sensor connected to the raspberry pi

Object detection and Optical character recognition

TensorFlow is a free and open-source software library for machine learning and artificial intelligence.We are using the tensor flow libraries for obstacle detection and optical character recognition.As we are using the Raspberry pi3 to process these deep neural network model results have shown that tensfor flow lite object detection model works much more efficiently and faster than the yolo algorithm.

In our project we have done the real time object detection using mobileNET_SSD algorithm from the TensorFlow lite framework. The TensorFlow lite is pre-trained on the COCO dataset.When the camera detects the object the class outputs which are detected are successfully converted into real time speech signal using the gTTS package of google in python library. So at this point our Artificial Eye model is able to detect any object with in the COCO dataset and convert it into an Audio format for the blind person to hear. As of now The coco dataset are trained only on 80 classes, for our future works we are trying to expand the dataset and train our own model to make it more accurate and effecient .

We will expand our model by customizing our own data for facial recognition of known people and gender recognition. We have also planned to implement the action detection model to convey the blind person about the action being performed around his environment.

MYCROFT

Mycroft is a additional feature which we are implementing to make our project a business model. The blind people can't see television ,ther only source of entertainment in audio. Mycroft is a open source,Artificial Intelligent voice assistant which can be easily integrated with a Raspberry pi. The blind person can easily interact with this model through his day asking about the weather, date ,time , hear some local news or listen to a music of his choice

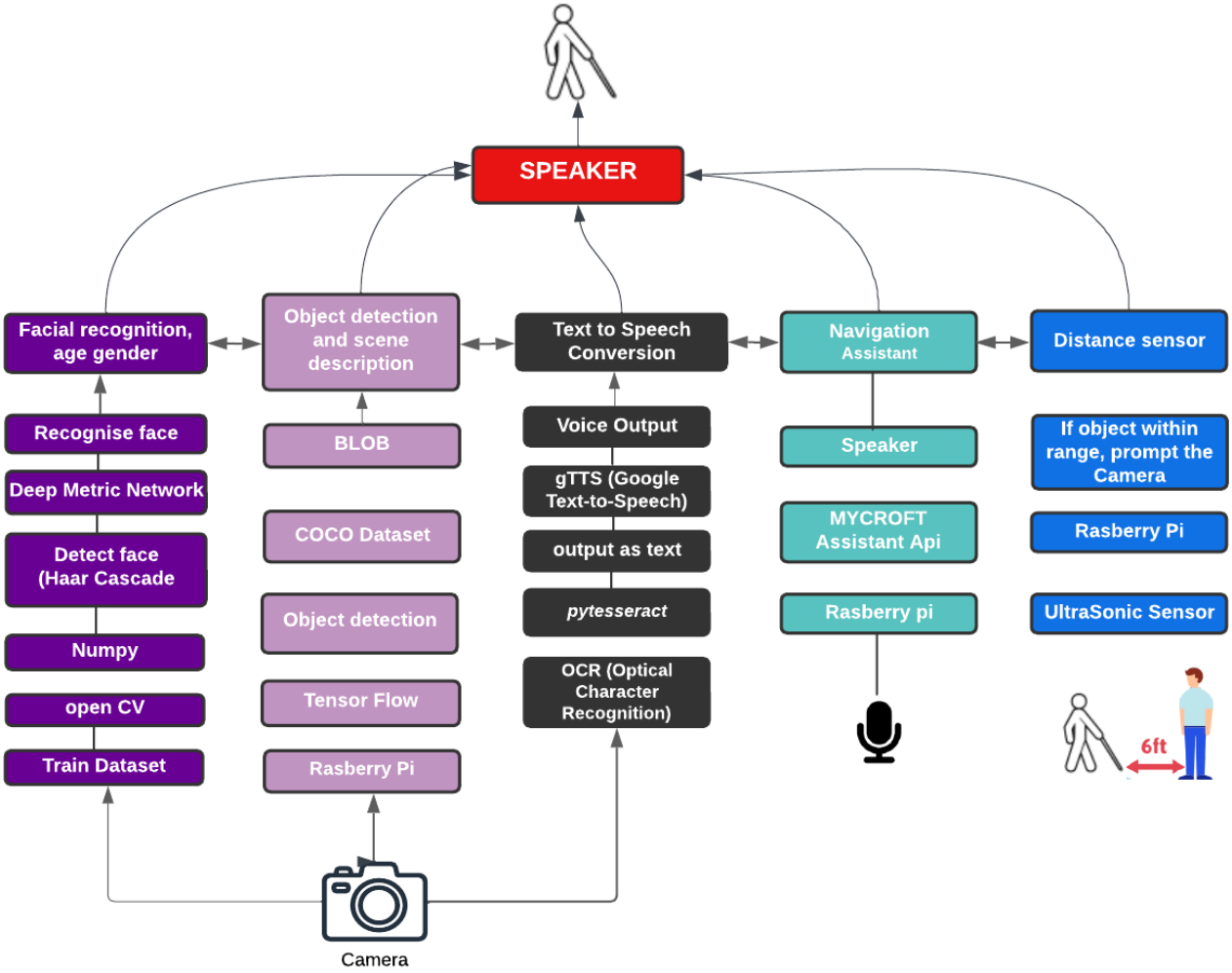


Fig 4: Tech Stack of the Artificial eye model

Fig 5 : The image of the Smart eye

Chapter 4 – Evaluation Results

Results drawn from the Proximity Sensor

The time taken by the raspberry pi to execute the Ultra sonic sensor are as follows

Distance(cm)	Time to execute
8.7	0.0037789
32	0.003605
61	0.00524
62	0.00587
65	0.00560188
77	0.00597
97	0.007149
118	0.00804
142	0.01002
161	0.0161

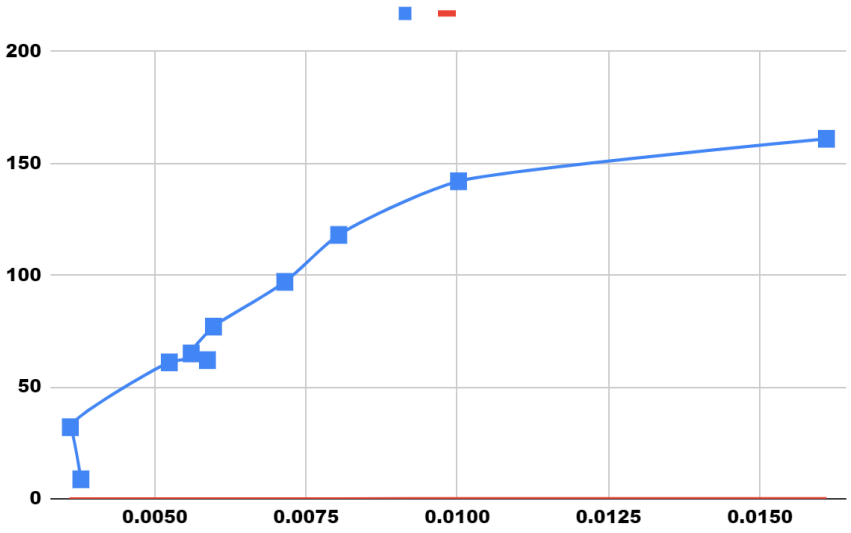


Fig 6: Graph plotting distance vs time take to execute by ultrasonic sensor

The Average response time of a human being is about 200milli seconds i.e upto 0.2 seconds but our ultrasonic sensor works much faster than the human eye with an average response time of 0.007137478 seconds or 0.7 milliseconds.
We can clearly say that our obstacle detection model is precise and accurate.

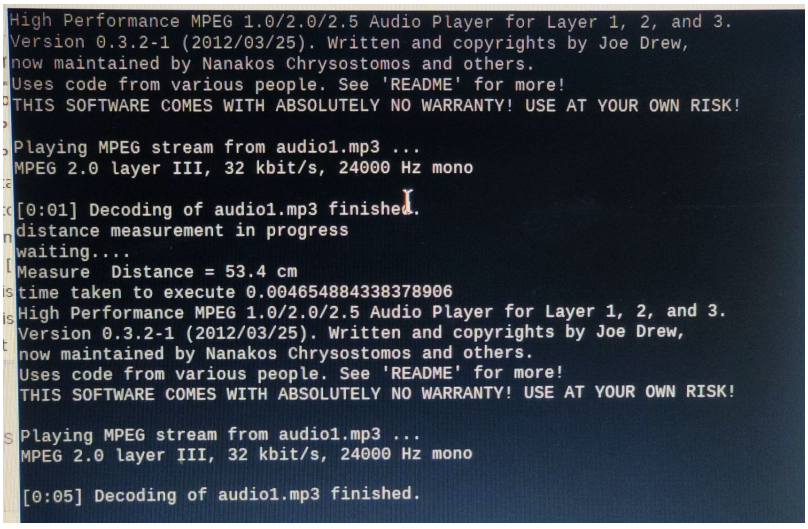


Fig7 : output of Ultrasonic sensor

Comparative analysis of the Object Detection model

Case 1: Implementing object detection on Rasberry Pi

Here we have done a complete performance analysis among most of the deep neutral network models for object detection. Deep neural network models like faster_RCNN , mobilenet-ssd,yolo v2, yolo v3,yolo v4 ,yolo v5 have been compared. We have compared these algorithms not only based on the mAP(mean Average Precision) but also GFLOPS (Giga Floating Point Operations Per Second)and MPARAMS (model parameters).
Most of the them only look for the results of only mAP but in this case as we are implementing this on raspberry pi we also have to consider to the computing speed of the it for which it is necessary to see the GFLOPS parameter.

What is GFLOPS?
GFLOPS stands for Giga_Floating_Point_Operations_Per_Second . It is a unit of measurementthat measures the performance of a floating point unit of a computer .Gigaflops measure the number of billions of floating-point calculations a processor can perform per second, and it directly tells us the computing power of a processor.
As we are using a raspberry pi which does not have a powerful cpu to handle the processing of the top notch CNN models like YOLO v4 , we have to consider a balance between with mAP and GFLOPS in such a way that we neither compromise too much on the accuracy by using lesser GFLOPS nor select an Algorithm with high accuracy with higher GFLOPS .

Hence we have made a comparative analysis of the following models and come to a conclusion that MobileNet_SSD works the best for us.

	GFlops	MParams	mAP	Source framework
faster_rcnn_inception_resnet_v2_	30.687	13.307	52.4	TensorFlow*
faster_rcnn_inception_v2_coco	30.687	13.307	40	TensorFlow*
Faster R-CNN Resnet-50	57.203	29.162	42.87	TensorFlow*
mobilenet-ssd	2.316	5.783	79.8377	Tensorflow*
ssd_mobilenet_v1_coco	2.494	6.807	23.3212	TensorFlow*
YOLO v2 Tiny	5.424	11.229	27.34	Keras*
YOLO v2	63.03	50.95	27.34	Keras*

YOLO v3 Tiny	5.582	8.848	35.9	Keras*
YOLO v4	128.608	64.33	71.17	Keras*
YOLO v3	65.984	61.922	62.27%	Keras*
YOLO V5s	17	7.3	55.4	Keras*
YOLO V5lite	2.42	1.62	41.3	Keras*

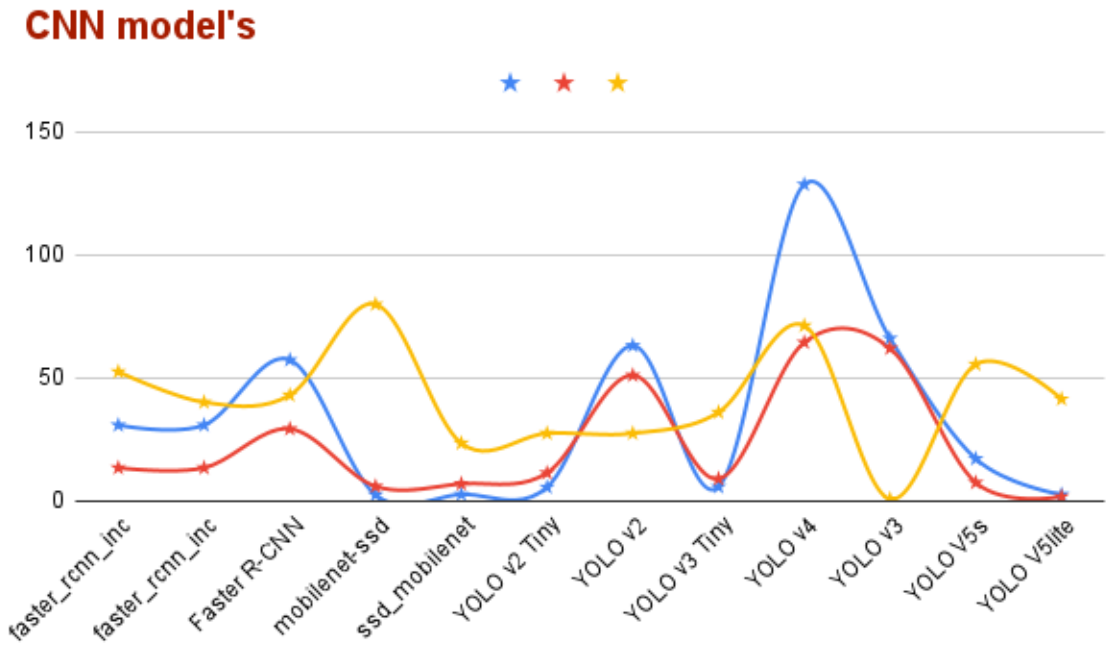


Fig 8 : Graph Ploting the Gflops,Mparams,mAP of different CNN models

Hence From this plotted graph it is very clear that the mobilenet_SSD uses the least GFLOPS with maximum mAP .Therefore we have used the TensorFlow Lite Framework which uses the mobileNet_SSD to achieve our object detection on the raspberry Pi

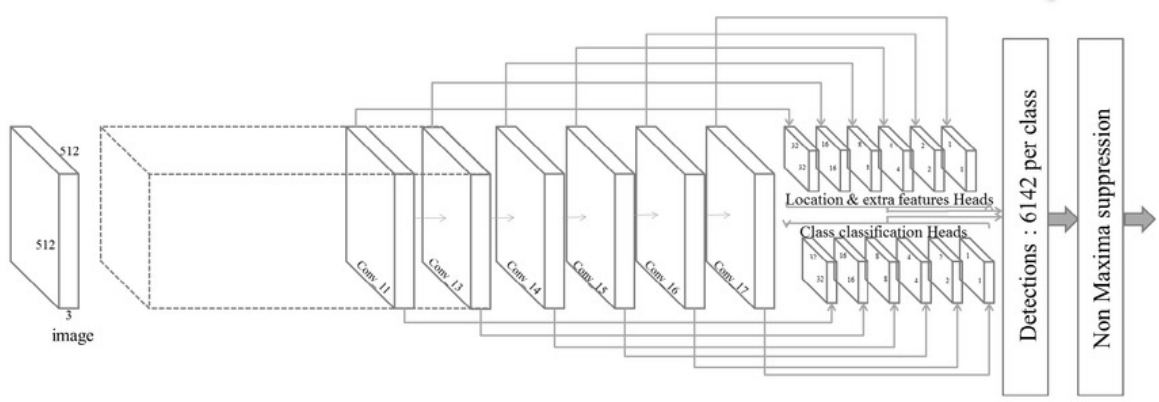


Fig 9: pictorial representation of MobileNet V1 based SSD architecture patter

CASE 2 : Implementing Object detection on powerful CPU's

Here, in this case we compare YOLO v5, YOLO v4 and YOLO v3's performance with respect to accuracy, precision, error, speed and many other aspects.With regard to this project RASPBERRY PI has been used and accordingly YOLO v5 was an efficient algorithm which was light weight with size only 1.7M (int8) and 3.3M (fp16). It can reach 10+ FPS on the Raspberry Pi 3B when the input size is 320×320 Bytes.

COMPARISON OF RESULTS:

ID	Model	Input_size	Flops	Params	Size (M)	Map@0.5	Map@.5:0.95
001	yolo-fastest	320×320	0.25G	0.35M	1.4	24.4	-
002	nanodet-m	320×320	0.72G	0.95M	1.8	-	20.6
003	yolo-fastest-xl	320×320	0.72G	0.92M	3.5	34.3	-
004	yolov5-lite	320×320	1.43G	1.62M	3.3	36.2	20.8
005	yolov3-tiny	416×416	6.96G	6.06M	23.0	33.1	16.6
006	yolov4-tiny	416×416	5.62G	8.86M	33.7	40.2	21.7
007	nanodet-m	416×416	1.2G	0.95M	1.8	-	23.5
008	yolov5-lite	416×416	2.42G	1.62M	3.3	41.3	24.4
009	yolov5-lite	640×640	2.42G	1.62M	3.3	45.7	27.1
010	yolov5s	640×640	17.0G	7.3M	14.2	55.4	36.7

Fig 10

mAP stands for mean average precision - mAP@0.5 means to say that the mAP calculated at IOU threshold 0.5.
Map@.5:0.95 means average mAP over different IoU thresholds, from 0.5 to 0.95, step of 0.05

What is the IOU threshold?
Intersection over Union, a value used in object detection to measure the overlap of a predicted versus actual bounding box for an object. The closer the predicted bounding box values are to the actual bounding box values the greater the intersection, and the greater the IoU value, on the other hand, If the difference between the predicted and actual bounding box is less, then lesser is IOU.

COMPARISON ON DIFFERENT PLATFORMS:

Equipment	Computing backend	System	Framework	Input	Speed{our}	Speed{yolov5s}
Inter	@i5-10210U	window(x86)	640×640	torch-cpu	112ms	179ms
Nvidia	@RTX 2080Ti	Linux(x86)	640×640	torch-gpu	11ms	13ms
Raspberrypi 4B	@ARM Cortex-A72	Linux(arm64)	320×320	ncnn	97ms	371ms

Fig 11

WHY IS YOLOv5 RECOGNISED AS THE BEST ALGORITHM?

Inference speed:The inference speed is measured with frames per second (FPS), namely the average iterations per second, which can show how fast the model can handle an input. The higher the inference speed, the faster, the better is the performance.YOLOv5 has higher inference speed which is why it is better than other YOLO models.

Detection of small or far away objects: Other models such as YOLO v3,v4 are not good when it comes to the accuracy of objects that are far away and of the objects that are small in size. YOLO v5 has a comparatively good accuracy when it comes to detecting objects in such a category.

Little to no overlapping boxes: YOLOv5 has great performance when it comes to separating bounding boxes for various objects at different distances apart from each other. Various other algorithms have a lot of overlapping issues with regard to bounding boxes which makes it difficult for detection of objects that are very near to each other.

Speed: This is a major advantage with the YOLOv5 model since it's an engine they developed to run sparse models on CPU. This is faster when compared to other GPU's and respective YOLO models.

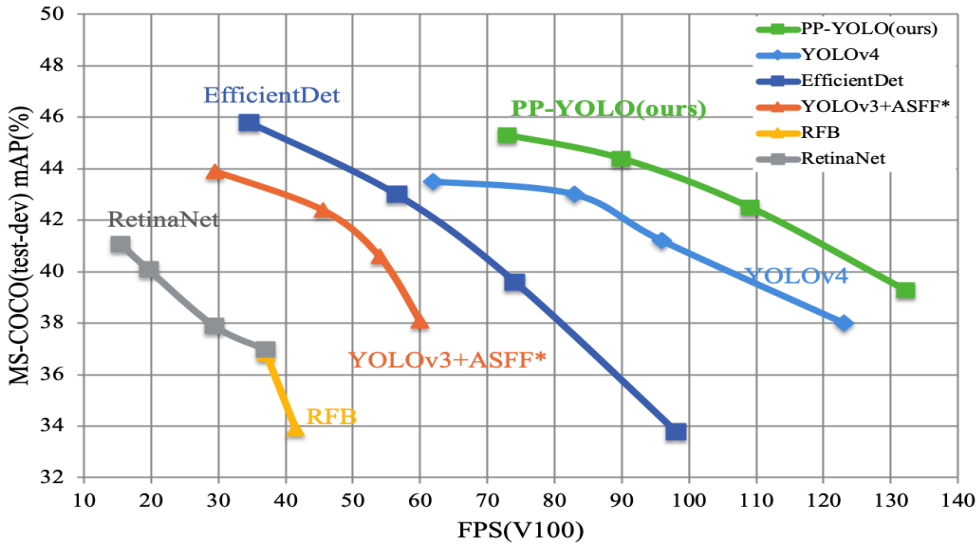

PP-YOLO(ours)
YOLOv4
EfficientDet
YOLOv3+ASFF*
RFB
RetinaNet
EfficientDet
PP-YOLO(ours)
YOLOv4
RetinaNet
YOLOv3+ASFF*
RFB
MS-COCO(test-dev) mAP(%)
FPS(V100)
50
48
46
44
42
40
38
36
34
32
10
20
30
40
50
60
70
80
90
100
110
120
130
140

Fig 12

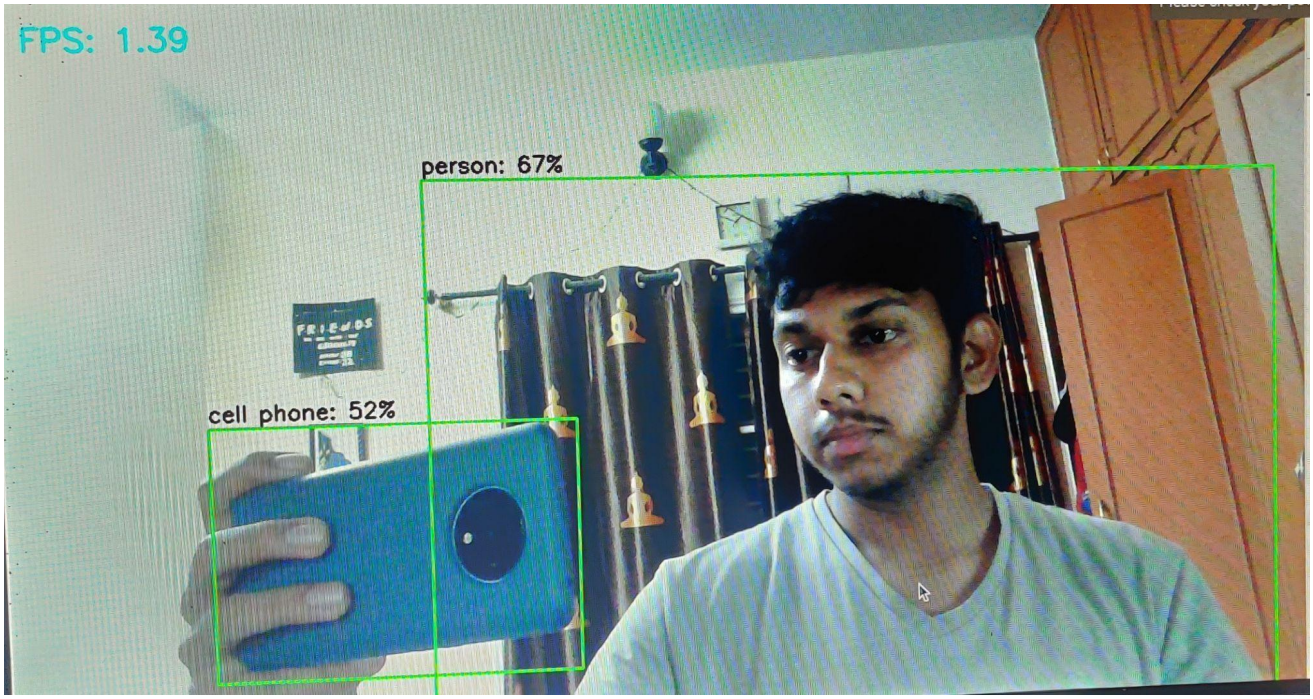

FPS: 1.39
person: 67%
cell phone: 52%

Fig 13

OCR ENGINE COMPARISON–TESSERACT VS EASY OCR:

1. ACCURACY: We have tested with texts and numbers of about 1000 samples. Generated random alphabets/numbers on a blank image , Tesseract and EasyOCR are used in parsing the image.

Scenario 1 : For 1000 sample images of alphabets with two words generated by random words and given input data flood experience

Scenario 2: For 1000 sample images of numbers with 5-digit and 2 decimal points numbers are used given input data 49403.65.

The detailed comparison and errors noticed:

	Error Rate on Numbers	Error Rate on Alphabets	Misinterpret Numbers	Misinterpret Alphabets
Tesseract	5.50%	0.70%	miss continuous 7 miss continuous 2 miss .	misinterpret t to r miss continuous l add ‘ in front of o add . in the end add , in the end
EasyOCR	1.90%	4.30%	miss 1 in the end misinterpret . to _	misinterpret l to i misinterpret h to n misinterpret f to t misinterpret d to a miss y in the end miss v

Fig 14

2. SPEED:for the same scenarios, below are the detailed comparisons

	CPU	GPU
Tesseract	0.3 seconds/image	0.25 seconds/image
EasyOCR	0.82 seconds/image	0.07 seconds/image

Fig 15

Conclusion:

- In conclusion, tesseract does a better job in recognising alphabets and easy OCRr's for recognition of digits .
- Tesseract is more recommended on a cpu and easy ocr on a gpu as it's more quicker .

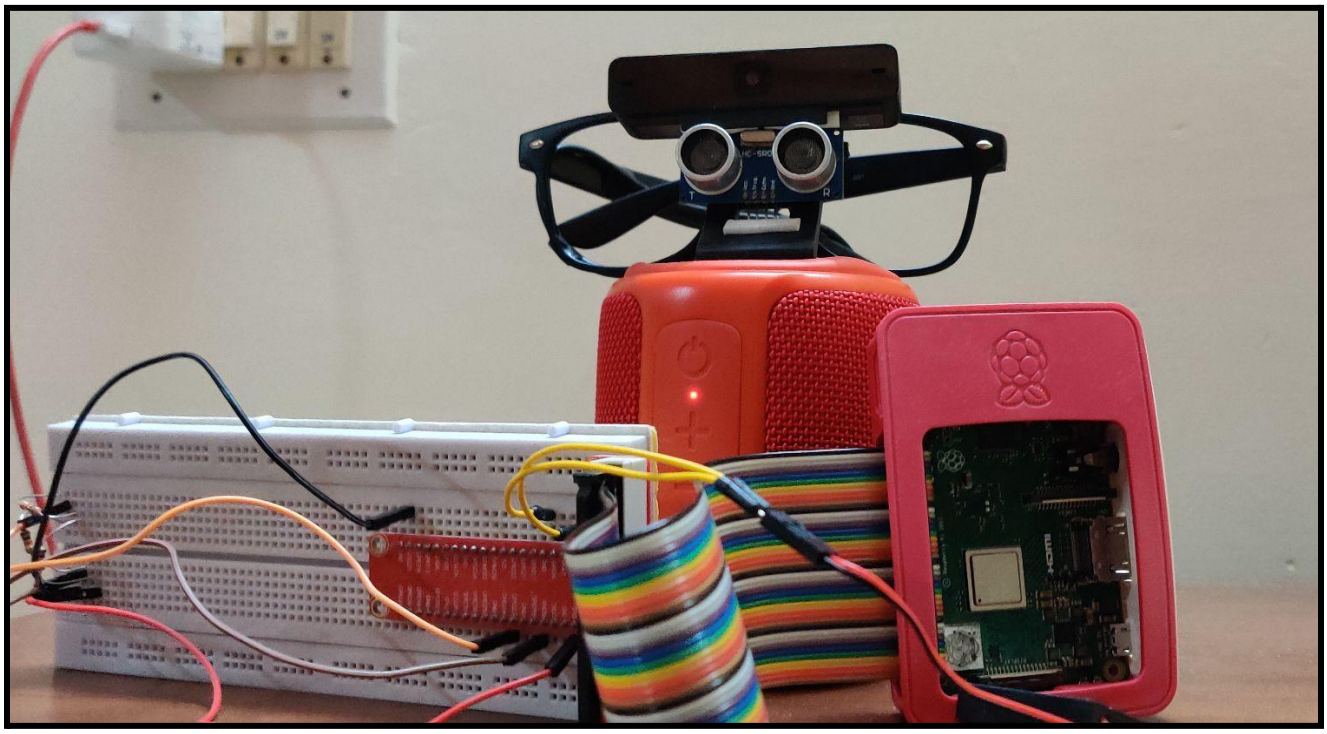


Fig 16

Chapter 5 – Conclusion and Future work

We have put together a functional prototype of a blind stick that helps the specially disabled to travel to the places of their choice with the help of multiple services offered by our model. By incorporating models that perform proximity detection using an ultrasonic SR04 sensor, object detection using a camera and finally OCR to read the label on the recognised object. Once read, the mycroft voice assistant will read out the recognised object via a speaker which provides guidance for the user about his surroundings for that instant. Based on some preliminary tests we found the average computing time of the whole sequence of processes to be 3-5 seconds.

Here are some scopes to our work in the future :

1. Involving facial recognition technique would help the person to recognise that someone who he knows is in the surroundings.
2. When the person is stationary and has a complete description of the surroundings, we could incorporate a unique process that helps the person in doing so.
3. Emotional intelligence too is a possible scope to explore as it gives the user an idea of the person's emotion.
4. Personal real time navigation could also be a productive feature that would basically behave like google maps but with additional benefits of the pre existing features.
5. Certain VQA models could also be constructed based on the current work which inturn helps the blind to ask queries regarding the surroundings and the objects he comes across in his daily life.
6. Lie detection is another intriguing domain which the model could assimilate. This would help the user to be not exploited in his life.
7. Better processors and investments are required to be put in to accommodate all the functionalities to run hand in hand and produce the best accurate result in minimal time.
8. Various models of IOT can be implemented parallelly.
9. Embedding the GPS functionality to track the person by his loved ones.

In conclusion, our work helps blind people to get a new and wider perspective of things and their surroundings in their day to day life, making them feel more empowered and confident.

References

[1] Jamuna S. Murthy,[1] G. M. Siddesh,[2] Wen-Cheng Lai ,[3,4] B. D. Parameshachari ,[5] Sujata N. Patil,[6] and K. L. Hemalatha[7], ObjectDetect: A Real-Time Object Detection Framework for Advanced Driver Assistance Systems Using YOLOv5

[2] T. V. Janahiraman and M. S. M. Subuhan, "Traffic Light Detection Using Tensorflow Object Detection Framework," *2019 IEEE 9th International Conference on System Engineering and Technology (ICSET)*, 2019, pp. 108-113, doi: 10.1109/ICSEngT.2019.8906486.

[3] Asifullah Khan,Anabia Sohail , Umme Zahoora , and Aqsa Saeed Qureshi Pattern Recognition Lab, DCIS, PIEAS, Nilore, Islamabad 45650, Pakistan Deep Learning Lab, Center for Mathematical Sciences, PIEAS, Nilore, Islamabad 45650, Pakistan)

[4] Joseph Redmon,Santosh Divvala, Ross Girshick , Ali Farhadi University of Washington , Allen Institute for AI , Facebook AI Research, You Only Look Once: Unified, Real-Time Object Detection

[5]C. L. Su, W. C. Lai, Y. K. Zhang, T. J. Guo, Y. J. Hung, and H. C.Chen, "Artificial intelligence design on embedded board with edge computing for vehicle applications," in IEEE 3rd International Conf. on Artificial Intelligence and Knowledge Engineering (AIKE), Laguna Hills, CA, USA, 2020.

[6] N. Dawar and N. Kehtarnavaz, "Action Detection and Recognition in Continuous Action Streams by Deep Learning-Based Sensing Fusion," in IEEE Sensors Journal, vol. 18, no. 23, pp. 9660-9668, 1 Dec.1, 2018, doi: 10.1109/JSEN.2018.2872862.

[7] Nisha Pawar, Zainab Shaikh, Poonam Shinde, Prof. Y.P. Warke ,Image to Text Conversion Using Tesseract Dept. of Computer Engineering, Marathwada Mitra Mandal's Institute of Technology, Maharashtra, India

[8] C. Lee, H. J. Kim, and K. W. Oh, "Comparison of faster RCNN models for object detection," in In 2016 16th international conference on control, automation and systems (iccas),pp. 107–110, Gyeongju, Korea (South), 2016.

[9] Veena N V, Chippy Maria Atony Msc Scholar, Assistant Professor Age and Gender detection using deep learning

[10] Smith, J. W., and Merali, Z. *Optical Character Recognition: The Technology and its Application in Information Units and Libraries*. The British Library,1995.

applied system innovation

MDPI

Article

Learning at Your Fingertips: An Innovative IoT-Based AI-Powered Braille Learning System

Ghazanfar Latif [1,*], Ghassen Ben Brahim [1], Sherif E. Abdelhamid [2,*], Runna Alghazo [3], Ghadah Alhabib [1] and Khalid Alnujaidi [1]

[1] Department of Computer Science, Prince Mohammad Bin Fahd University, Khobar 34754, Saudi Arabia; gbrahim@pmu.edu.sa (G.B.B.); 202000093@pmu.edu.sa (G.A.); 202002530@pmu.edu.sa (K.A.)
[2] Department of Computer and Information Sciences, Virginia Military Institute, Lexington, VA 24450, USA
[3] Department of Education, Health, & Behavioral Studies (EHBS), University of North Dakota, Grand Forks, ND 58202, USA; runna.alghazo@und.edu
* Correspondence: glatif@pmu.edu.sa (G.L.); abdelhamidse@vmi.edu (S.E.A.)

Abstract: Visual impairment should not hinder an individual from achieving their aspirations, nor should it be a hindrance to their contributions to society. The age in which persons with disabilities were treated unfairly is long gone, and individuals with disabilities are productive members of society nowadays, especially when they receive the right education and are given the right tools to succeed. Thus, it is imperative to integrate the latest technologies into devices and software that could assist persons with disabilities. The Internet of Things (IoT), artificial intelligence (AI), and Deep Learning (ML)/deep learning (DL) are technologies that have gained momentum over the past decade and could be integrated to assist persons with disabilities—visually impaired individuals. In this paper, we propose an IoT-based system that can fit on the ring finger and can simulate the real-life experience of a visually impaired person. The system can learn and translate Arabic and English braille into audio using deep learning techniques enhanced with transfer learning. The system is developed to assist both visually impaired individuals and their family members in learning braille through the use of the ring-based device, which captures a braille image using an embedded camera, recognizes it, and translates it into audio. The recognition of the captured braille image is achieved through a transfer learning-based Convolutional Neural Network (CNN).

Keywords: braille learning; education for the blind; internet of things (IoT); deep learning; smart device; convolutional neural networks; transfer learning

Citation: Latif, G.; Brahim, G.B.; Abdelhamid, S.E.; Alghazo, R.; Alhabib, G.; Alnujaidi, K. Learning at Your Fingertips: An Innovative IoT-Based AI-Powered Braille Learning System. *Appl. Syst. Innov.* **2023**, *6*, 91. https://doi.org/10.3390/asi6050091

Academic Editor: Dorota Temple

Received: 16 July 2023
Revised: 11 September 2023
Accepted: 28 September 2023
Published: 11 October 2023

1. Introduction

Braille is the universal form of literacy for the blind and visually impaired. Braille bridges the communication gap between the visually impaired and their surroundings. It is the only textual representation that the blind and the visually impaired can understand. A major challenge faced by the blind and the visually impaired is their need to learn the braille language to be able to read and learn in general. They require a dedicated instructor and one-to-one supervision to learn braille. Visually impaired individuals living in small cities or rural areas find it challenging to learn the language due to the limited number of educational institutes and special education schools present in these locations. There is also a shortage of resources provided for enhancing the learning environment of these individuals. Furthermore, the process of learning braille is both time-consuming and requires specialized personnel due to the need for a specialized instructor to guide and assist the blind in learning. Therefore, the aforementioned challenges cause many visually impaired individuals to feel discouraged and unmotivated to learn. If an automated translation system exists that is proficient in braille, then it will accelerate the learning process for the visually impaired because computers can process information and translate braille text much faster than humans [1]. This means that a visually impaired individual

would be able to read a braille text much faster by having the automated IoT-based system directly translate a braille document to audio.

Braille recognition systems translate printed braille to its textual and natural language representations. The dotted format of braille documents is the starting point for any automated recognition system. The system should be able to capture the dotted format of the braille language, translate it to the corresponding alphabet letter, and combine the letters to recognize words. There are different braille systems for different languages; thus, the development of an automated braille translating system should target a particular language or should at least have the capability to choose between languages for bilingual users. The Automatic braille recognition system could use computer vision and machine learning models like Conventional Neural Networks (CNN), Decision Trees (DT), K-nearest neighbor (KNN), and Support Vector Machines (SVM) for this purpose. Deep learning (DL) and computer vision are extensively used for pattern recognition and image classification [2–4].

The main goal of this work is to develop an automated system based on artificial intelligence for Arabic and English bilingual individuals who are visually impaired. The main reason for the exact choice of these two languages is that English is taught in many Arabic-speaking countries; thus, most individuals in Arabic-speaking countries—especially in the Middle East—are bilingual (proficient in Arabic and beginner to intermediate in English). Thus, the goal is to develop a device that will assist in the following ways:

1. Teach visually impaired individuals braille with a learn-at-your-own-pace methodology without the need for professional Braille instructors;
2. Teach braille to the parents of visually impaired individuals so that they can in turn teach braille to their children;
3. In terms of recognizing braille characters, assist visually impaired individuals in reading braille documents and books at much faster speeds and with a high accuracy level.

This research work offers a 4-fold contribution consisting of these objectives:

1. Presents an extensive survey of existing techniques to detect Braille in different languages;
2. Designs an IoT-based system that can fit on the ring finger, simulating the real-life experience of a visually impaired person;
3. Develops an ML-based model to recognize and translate Arabic and English braille into audio using deep learning techniques with transfer learning;
4. Creates a new bilingual Arabic–English braille dataset, which is to be expanded using data augmentation techniques;
5. Perform a performance evaluation study of the entire system with regard to accuracy and effectiveness.

The research topic is significant because the visually impaired lack access to both educational centers that have Braille translation systems and instructors for the learning process. It is estimated by the INEI that only 23.9% of visually impaired individuals manage to complete their education, thus indicating the need for a system that supports translation from Braille to text for the integration of the visually impaired into their communities. The implementation of language translation systems is crucial to restrict the communication gap, and performing further research is important for providing open sources on how to build translators. Sometimes, a person may wish to learn braille to teach it or to communicate with someone with visual disabilities. This improves the daily life activities of the visually impaired [5].

The rest of the paper is organized as follows: Section 2 discusses a review of the recent studies, Section 3 explains the methodology proposed, the experimental results are discussed in Section 4, and the research work is concluded in Section 5.

2. Review of Recent Studies

In [1], researchers suggest a deep learning scheme for character detection with a position-free touchscreen-based input methodology. This device translates braille input into

natural language by simply tapping on the dots of each character. The dataset used in this research is composed of 1258 photographs of sizes 64 × 64 with two categories: Category-A (a–m) and Category-B (n–z). The dataset was obtained from a screen interface for Android devices. The input braille text is processed and entered into the Convolution Neural Network (CNN). Two CNN techniques were used: transfer learning and the sequential model. The recognition is achieved using a deep learning model trained using the gathered braille dataset. The classification evaluation was carried out using DL techniques such as the GoogleNet Inception model, achieving an accuracy of 95.8%, and the sequential model, achieving a total accuracy of 92.21%.

In [6], the authors proposed a touchscreen to detect Urdu braille characters using ML methods. The dataset obtained from the National Special Education School is composed of 39 classes sorted into three groups with 13 classes in each group, 144 cases for each class resulting in 5616 cases in total. The letters are input into the screen. The methodology uses a Reconstruction Independent Component Analysis (RICA)-based feature extraction model. The highest-achieving classifier was the support vector machine (SVM) with a yielded accuracy of 99.73% accuracy. However, other robust ML techniques were used such as K-nearest neighbors (KNN) and decision trees (DT) for comparison purposes. The evaluation was conducted in terms of total accuracy, true positive rate, true negative rate, false positive rate, positive predictive value, negative predictive value, and area under the receiver operating curve. Unfortunately, this study is only limited to Grade 1 Urdu braille and does not include Grade 2 Urdu braille with speech and text responses.

In [7], the authors suggest using RICA-based feature extraction methods and automated tools to extract English braille alphabets. The proposed methodology uses a Grade 1 English braille dataset obtained from a touchscreen from the National Special Education School along with a position-free braille text entry technique to produce synthetic data to generate a dataset composed of 2512 cases. The dataset comprises 26 braille English letters and is divided into two classes: class 1 (1–13) and class 2 (14–26). For character recognition, Decision Trees (DT), Support Vector Machine (SVM), and K-nearest neighbor (KNN) with PCA-based feature extraction methods and Reconstruction Independent Component Analysis (RICA) were implemented. RICA outperformed PCA and the SVM classifier also achieved an accuracy of 99.85%. Sequential methods and RF methods yielded the highest accuracy with a value of 90.01%. The performance was evaluated based on total accuracy, true positive rate, true negative rate, false positive rate, positive predictive value, negative predictive value, and area under the receiver operating curve. The accuracy achieved is 100% for classes such as a, c, d, h, i, j, p, u, w, and k, 99.87 and 99.60% for other classes such as b, f, q, s, t, and v. The study is only suitable to Grade 1 English character braille and cannot be implemented with restricted computation power. The study also does not use DL methods such as CNN and GoogleNet to enhance the outcome.

Authors in [8] recommend using a Histogram of Oriented Gradient Features and a Support-Vector Machine (SVM) for braille recognition and feature extraction. The method can translate Sinhala braille to Sinhala language and English braille to the English language. The images are processed, segmented, and then recognized using HOG feature extraction methods and the SVM classifier method. The study uses two types of HOG feature extraction methods: a cell size of 4 × 4, and another one of 2 × 2. The dataset is composed of both scanned handwritten and computer-generated braille text. The methodology can process Grade 1 English characters as well as some Grade 2 characters. The yielded accuracy was 99%. The authors report that higher processing time was needed in the case of 2 × 2 cells compared to 4 × 4 cells.

Reference [9] advocates for using a Semantic Retrieval System to assist visually impaired individuals in mathematical studies. The methodology begins with translating a query math formula in braille into MathML code, and then the structural and semantic meaning is obtained from the MathML expression to produce a multilevel tree. The feature extraction method used is the conventional vector model. Afterward, in the classification stage, the K-nearest neighbors method is used to choose a multilevel similarity measure to

compare between expressions. Lastly, the query produced is translated to braille mathematical expressions. The dataset was created using MathType and consists of 6925 mathematical equations and expressions from five languages: Hebrew, Japanese, Tifinagh, Arabic, and Latin. For each language, 1385 different types of equations were written. This study used Latin to test the performance of the methodology.

Authors in [10] have used a novel approach of the CNN extraction method to translate Bangla handwritten text to Bangla braille notation. The study used an object detection model, Faster-RCNN to draw boundaries over Bangla cells and then used 10 CNN models for classification. Faster-RCNN is a fast and efficient algorithm. The CNN models used include VGG16, DenseNet201, ResNet152V2, MobileNet, and ZFNet. The CNN models were trained and tested using the Microsoft Azure ML platform for calculation using Standard_NV48s_v3. Results show that the highest achieving accuracy CNN model was VGG16 with a value of 95%. The methodology was implemented using Python v3, Keras, and TensorFlow libraries. Furthermore, the dataset was collected from external resources of handwritten Bangla, the images were resized using a canny edge detection, and a median filter was applied to decrease the noise and threshold. Afterward, it is converted to black and white. The dataset comprises 105 classes with 157,500 photographs where 80% were kept for training and 20% were kept for testing. Each class comprises 1500 photographs. Unfortunately, this study covered a limited number of conjunctions and many of the 300 conjunctions of the Bangla language were not considered.

In [11], the paper recommends using machine learning (ML) for character recognition of Hindi handwritten documents to translate to braille text. The pages are first transformed into a printable form and then converted to braille using UTF-8 codes. The dataset used is composed of 92,000 images and for each of the 46 characters, 2000 images are used for classification. However, vowels and Matras are discarded from the dataset. Additionally, the author uses a Histogram of oriented gradient features of Hindi characters to extract features. The segmented letters are then classified using an SVM classifier for character recognition. To produce higher levels of accuracy, the resolution of the image should be greater than 300 dpi. Further, the range of accuracies achieved is between 87.667% to 97.667%. This study is unique because it tackles a language with limited resources. The results showed that the classifier failed to predict the letters "HA" and "DHA", which is considered a limitation of the proposed model. However, because the cell size used was upgraded to 4×4 the average accuracy increased from 94.65% to 95.56%.

In [12], the study encourages using a Convolution Neural Networks (CNN) system to classify images of braille and translate them to English characters. The dataset used is composed of 14,378 braille photographs. The 3-major steps in the conversion process are pre-processing segmentation and image classification. In pre-processing, the method uses grayscale conversion, contrast adjustment, finding circles, and inverting colors. Segmentation is divided into line segmentation and cell segmentation. For image classification, DL algorithms and CNNs are used with nine different layers including, an input layer, convolutional filter, max pooling, and output layer, etc. The paper yields a high accuracy with a value of 96.37% for a 500-image containing dataset size. Furthermore, the scheme possesses high-performance characteristics due to the implementation of deep learning and not only simple neural networks.

Authors in [13] consider a deep learning-based model that combines the CNN model to detect characters and transformer models to recognize words. The results showed that the proposed model achieves high performance in terms of accuracy in detecting characters and words reaching 98.6% and 96.7%, respectively.

In [14], the authors proposed a hardware device to aid visually impaired individuals. This device combines the use of long short-term memory (LSTM) along with Raspberry Pi and the convolutional neural network (CNN). The proposed system recognizes numbers, letters, dots, and punctuation. Performance-wise, the system achieved a high level of accuracy, reaching 98%.

Artificial intelligence (AI) and Deep Learning (ML) have been used in research to assist students with disabilities as well as in other fields such as the medical field, sign language, and handwritten text classification. In [15], the authors proposed an automated AI-based system for assisting the deaf and hard of hearing to communicate with their surrounding community. Using Random Forest (RF), the authors reported an accuracy of 92.15%. In [2], the author proposed a system for assisting the deaf and hard of hearing using deep learning (DL). They reported an accuracy of 97.6%. In [3], the authors proposed an automatic AI-based system for the automatic recognition of multi-lingual handwritten digits using novel structural features. They reported an accuracy of 96.15%. There are many more examples of AI and ML being used to automate and develop automatic systems in many fields including medicine, agriculture, education, etc. The continued pursuit of optimal solutions will develop over time until the optimal solutions are reached and developed into patented devices that could actually be used and assist in making people's lives better.

Attempts were also made to design a model to perform a reverse operation of what this current research aims. For instance, the authors in [16] designed a CNN-based model to recognize real-time Arabic speech and eventually translate it into Arabic text then convert it into Arabic braille characters. The model works on digits and is yet to be improved to include alphabets. An accuracy performance of 84% was achieved when adding the ReLU activation function to the CNN model.

3. Proposed Methodology

The proposed system shown in Figure 1 is designed to be compact, portable, and fitting on the tip of a finger. Equipped with a digital camera, it is capable of capturing images of the braille dots for processing. The dimensions of each braille dot are determined based on the tactile resolution of a person's fingertips. The dot's height measures approximately 0.5 mm (0.02 inches), with a vertical and horizontal spacing of 2.5 mm (0.1 inches) between dot centers and a spacing of 3.75 mm (0.15 inches) between adjacent cells. A standard braille document measures 11 $\times$ 11.5 inches with each line having between 40 and 43 cells.

Figure 1 shows a detailed workflow of the proposed system. During the AI software processing phase, the captured image with the help of a button will be segmented to exclude the region of the image that does not contain braille dots. The IoT system follows a series of image processing steps, including edge detection, binary conversion, hole fitting, and image filtering. Preprocessing methods are used to reduce noise and enhance the visibility of the dots. The system also performs segmentation to allow for individual identification of the letters. During the next step, the image is resized to 16 $\times$ 16 pixels. In the braille system, each letter is represented by a single cell consisting of six dots arranged in two columns and three rows. Once the image has been extracted, the system undergoes training to classify the braille characters based on their corresponding classes. Each letter or number is associated with a specific class, allowing for accurate mapping. The performance of the algorithms used to train the models is evaluated in terms of accuracy, positive and negative predicted values, and other relevant metrics. It is important to note that misclassification errors may arise due to challenges encountered during noise removal, variations in braille dot sizes, and the process of segmentation.

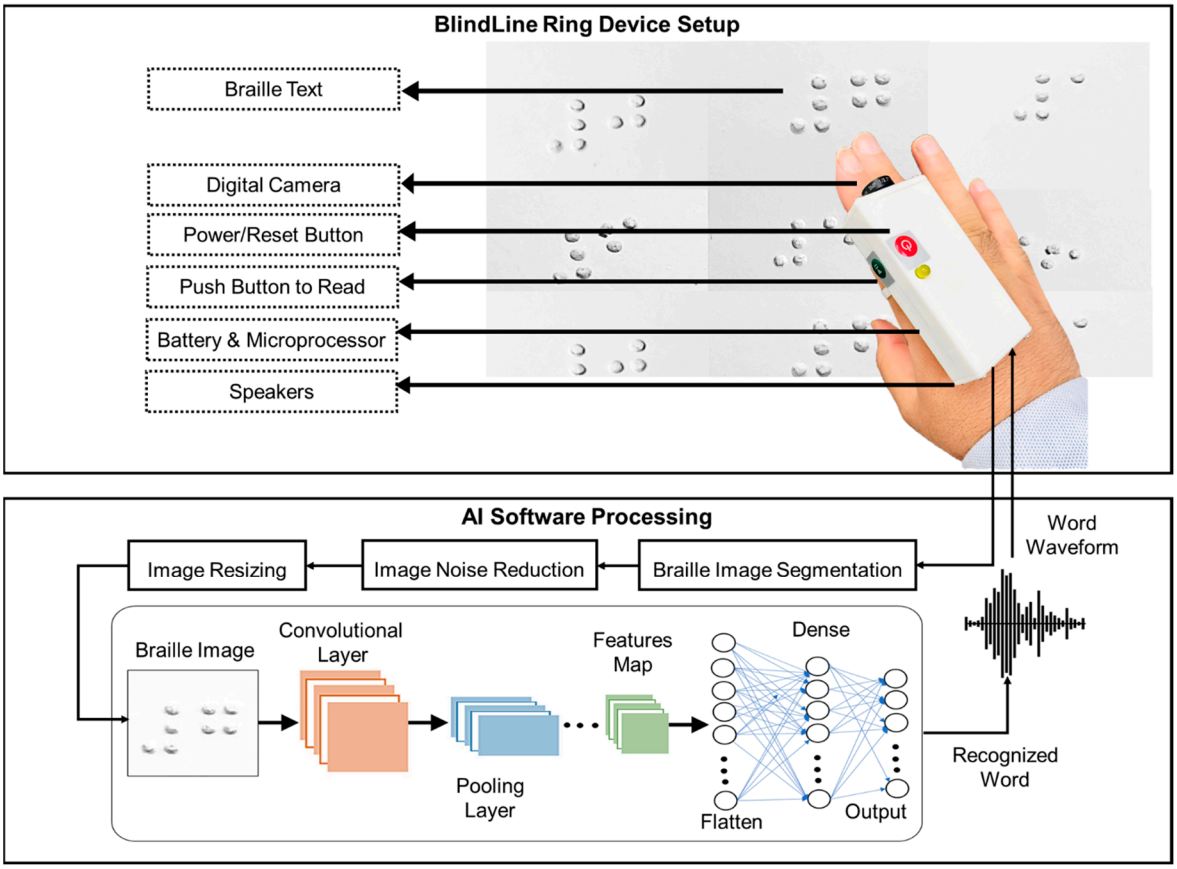


Figure 1. Workflow of the proposed system for the IoT-based braille language learning on finger tip.

3.1. Experimental Dataset

This research was conducted on a new built dataset containing images of Arabic and English braille characters. The dataset is used as an input to test the validity and efficiency of the proposed methodology and is composed of 28 Arabic characters (from "أ" to "ي"), and 26 English characters (from "a" to "z") as shown in Figures 2 and 3, respectively. The different augmentation methods are applied to the collected images including width-height shift, rotation, and brightness which change the shift, rotational, and brightness values, accordingly. English braille dataset is composed of 'A' to 'Z' English alphabetical letters and comprises 500 labeled images for each class which is deemed sufficient for the training, validation, and testing of the model for braille dots. Similarly, the 26 Arabic characters dataset was also augmented to have 500 labeled images of each character's class used for the training while another 15 non-augmented images of each character were used for testing. The images were cropped individual letters and the image name contains the number of the image, the character alphabet, and the type of data augmentation. The images in the dataset possessed different brightness for better machine learning training and character recognition. It is important to mention that the detection of braille characters may be challenging due to their small size, the minimized visual contrast with their background, similarity between characters. Our dataset design involved printing braille letters on single-sided A4 embossed paper in blue and white, creating the images. These images were captured using smartphone cameras, ensuring diversity by varying lighting conditions, colors, angles, and heights. To optimize processing, the images were converted to grayscale, and resized to 256 pixels.

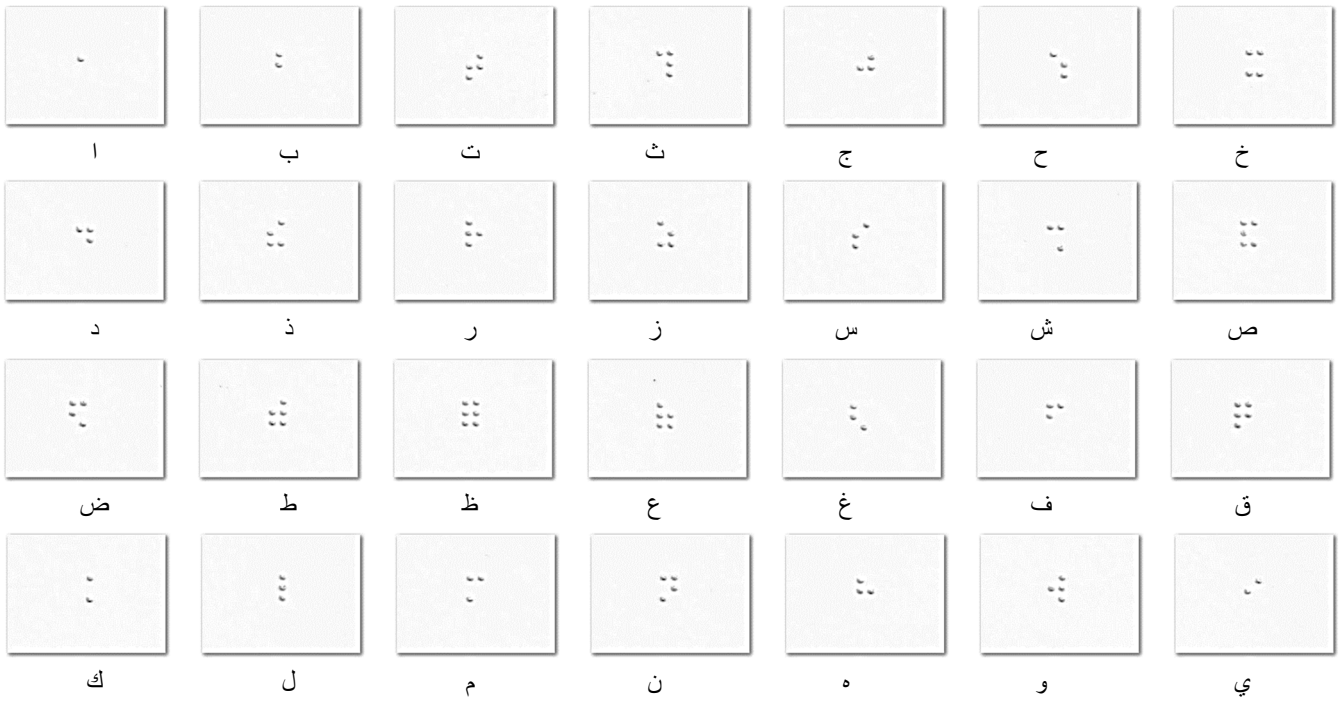


Figure 2. Sample braille for the Arabic braille language for Arabic alphabets.

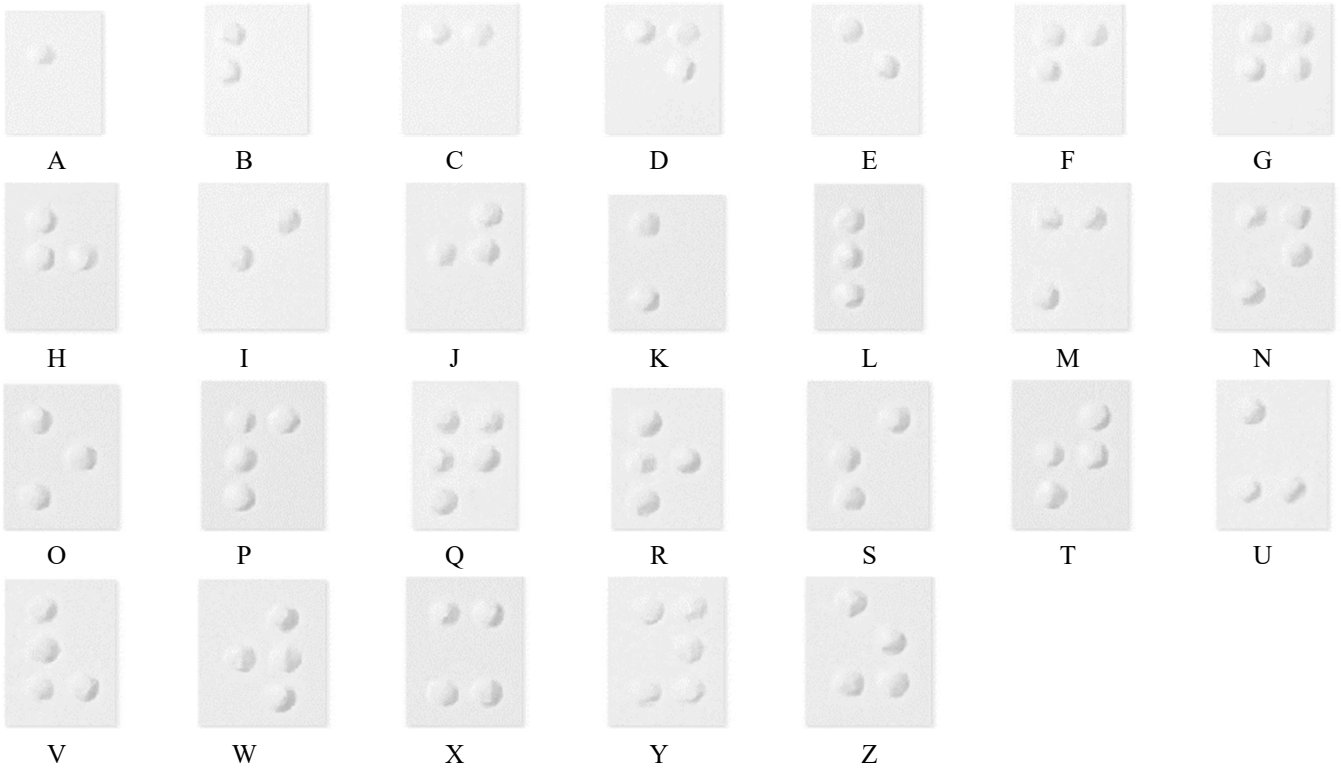


Figure 3. Sample braille for the Arabic braille language for English alphabets.

3.2. Convolutional Neural Network-Based Transfer Learning

Convolutional Neural Network (CNN) is an algorithm widely used in computer vision and deep learning. The algorithm takes an image as input and assigns significance to several objects in that image to distinguish one from the other. CNN algorithm requires minimal pre-processing compared to other classification methodologies. The CNN-based

models are generally divided into three major layers: the convolutional layer, the pooling layer, and the fully connected layer [17,18]. The algorithm begins with reducing the image into an easily processed form while preventing the loss of significant features. This aids the creation of an architecture that can learn features and is scalable to interpret new datasets.

In the convolutional layer, the Kernel/Filter, K, is the element performing the convolution operation in the first part of the layer. The filter traverses the image by moving to the right until it covers the full width and then down until it covers all pixels. The goal of convolution operations is to extract high-level features of an image. The results are of two types: dimensionality is either increased or stays the same by applying the same padding or convolved features reduced in dimensionality by applying valid padding. CNN may have multiple convolutional layers. The first layer captures the low-level features and with additional layers, it adapts the high-level features. This builds a system that can interpret images.

The next layer is the pooling layer, where the spatial size of the convolved feature is reduced to minimize the computational power necessary for data processing through dimensionality reduction. The pooling layer extracts rotational and positional invariant dominant features for model training. Pooling has two types: average pooling and max pooling. Average pooling computes the average of all the values from the section of the image covered by the kernel. On the contrary, max pooling selects the maximum value from the section covered by the kernel and implements a noise suppressant.

The convolutional layer and the pooling layer compose the ith layer of a CNN. Each architecture has a unique number of layers depending on the complexity of the image. Increasing the number of layers assists in capturing additional low-level details but requires more computational power. Now, the model can interpret the features and complete the first stage of the architecture to then move to the next stage and feed the classification model. In the third and last stage or the fully-connected layer, the image is flattened into a column vector and is fed into the neural network. The model then differentiates between significant and insignificant features and classifies them using the SoftMax classification technique. With each layer, the model increases in complexity and can identify more sections of a photo. Earlier layers extract simpler features and later ones extract more elements used to identify the object [19].

ConvNet includes several architectures such as LeNet, AlexNet, DenseNet, GoogleNet, and VGGNet [20]. These models are widely adopted as transfer learning to retrain the models with the new datasets for different applications. AlexNet is an extension of LeNet with a deeper architecture. It has eight layers in total: five convolutional layers and three fully connected layers. All layers are connected to a ReLU activation function. AlexNet employs data augmentation and dropout techniques to prevent overfitting due to excessive parameters.

DenseNet can be considered an extension of ResNet, where the output of a previous layer is added to a subsequent layer. DenseNet proposes concatenating the outputs of previous layers with subsequent layers, which enhances the distinction in the input of succeeding layers, thereby increasing efficiency. DenseNet significantly reduces the number of parameters in the learned model. For this research, the DenseNet-201 architecture was used. It has four dense blocks, each followed by a transition layer except for the last block, which is followed by a classification layer. A dense block contains several sets of 1×1 and 3×3 convolutional layers, while a transition block contains a 1×1 convolutional layer and a 2×2 average pooling layer. The classification layer in DenseNet-201 consists of a 7×7 global average pool followed by a fully connected network with 28 outputs based on the 28 Arabic braille letters.

GoogleNet architecture is based on inception modules, which perform convolution operations with different filter sizes at the same level. This increases the width of the network. The architecture has 27 layers (22 layers with parameters) and nine stacked inception modules. At the end of the inception modules, a fully connected layer with a SoftMax loss function serves as the classifier for the 28 classes of Arabic braille letters.

3.3. Fine-Tuned VGG16 Architecture

Large-scale visual data classification is usually performed using VGG16 and VGG19 CNN architectures. VGG16 is a CNN that could be combined with transfer learning for the classification process [21]. VGG16 is divided into three parts: convolutional layers which utilize filters for feature extraction from images, pooling layers for reducing spatial size, thereby decreasing the number of parameters and computations, and fully connected layers for final classification. When combining VGG16 with transfer learning, the model is expected to become more accurate, faster, and require less training time. This is a result of the fact that VGG16 is already pre-trained on large datasets and thus can detect particular features. Transfer learning allows leveraging the VGG16 pre-trained weights thereby increasing efficiency.

Small convolutional filters are used in the VGG16 architecture to increase network depth. The input is of size 224 × 224 × 3, where 3 refers to 3 color channels. As depicted in Figure 4, the input images go through the convolutional layers along with the small receptive field of size 3 × 3 and the max pooling layers. As shown in Figure 4, the first two sets of VGG utilize conv3-64 followed by a conv3-128 layer, using the ReLU activation function. The remaining three sets use conv3-256, conv3-512, and conv3-512, respectively, also utilizing the ReLU activation function. A stride of 2 and 2 × 2 always accompanies the convolutional layers in VGG16 and VGG19, while varying the number of channels between 64 to 512. It should be noted that the only difference between VGG19 and VGG16 is the presence of 16 convolutional layers. The fully connected layer usually has outputs representing the number of classes and in this case, it has 28 outputs corresponding to the 28 Arabic braille letters.

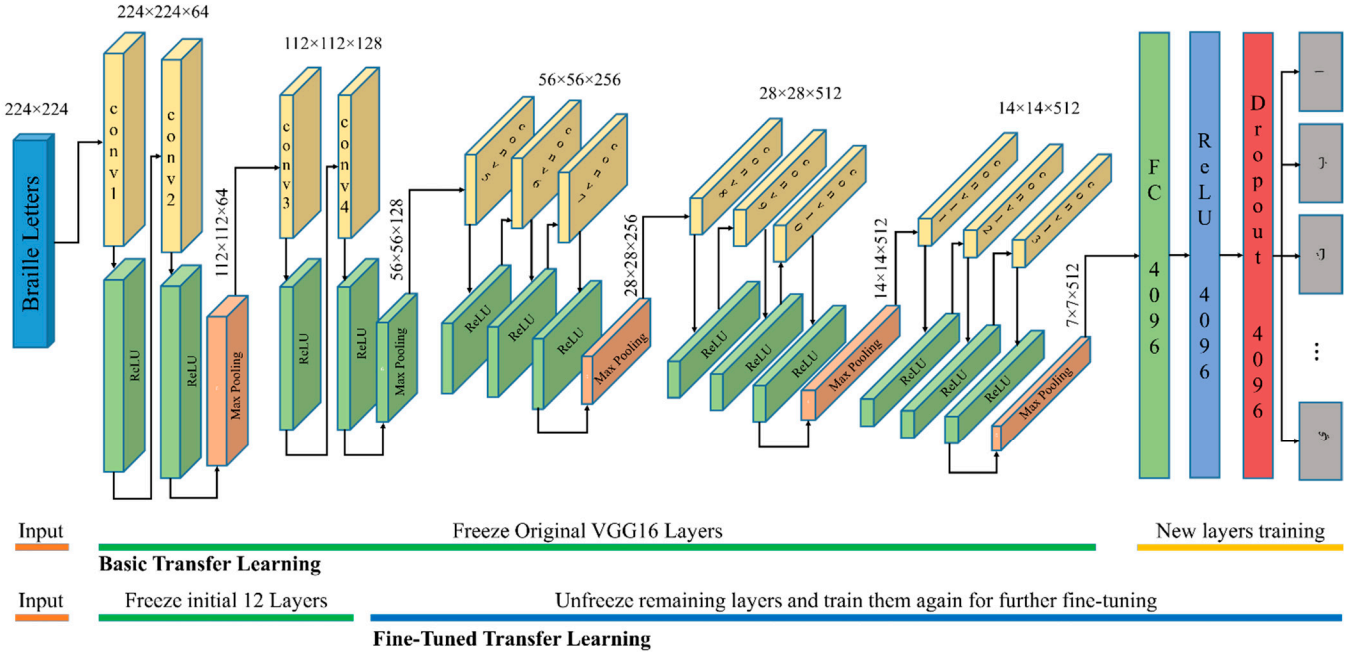


Figure 4. Fine-tuned VGG16 architectures for braille language detection.

4. Results and Discussions

The original and augmented datasets were used in the experiments in order to increase the overall size of the dataset. Various metrics were used in order to evaluate the performance of the proposed methodology. These include recall, precision, accuracy, and F1 measure [22]. In the proposed model, the idea is to freeze the top twelve layers and unfreeze the remaining layers to retrain the unfrozen layers. The decision to freeze the initial layers and retrain the later layers was made to balance pre-trained knowledge while adapting to our specific task. The determination of optimal layers for freezing and retraining was based on systematic experimentation, aiming for a balance between prior knowledge and

task-specific adaptation. This approach was applied to various deep learning models including VGG19, VGG16, DenseNet, AlexNet, GoogleNet, and LeNet. The proposed approach was applied to the combined dataset and to both the Arabic and English braille letters. In order to compare the performance of the proposed approach, the first experiment was performed using the original freeze weight of the original CNN models applied to the Arabic braille language dataset. The results are shown in Table 1. It should be noted that each letter has 500 images being used. These are divided into 300 images (60%) of each letter for training, 100 (20%) for validation, and 100 (20%) for testing. This percentage was used for all letters in both the Arabic and English braille alphabets. The experiments are performed for 30 epochs with a batch size of 512 with Adam optimizer and a learning rate of 0.001.

Table 1 shows the results of the experiments of the original CNN models using freeze weight applied to the Arabic braille language dataset. The results indicated that the best accuracy was achieved using GoogleNet with an average value of 98.63% and 98.4%, 98.4%, and 98.1% for precision, recall, and F1-measure, respectively. The lowest accuracy was reported for the GoogleNet algorithm with an average accuracy of 94.50%.

Table 1. Experimental results of freeze weights of the original CNN models for the Arabic braille language data.

	Accuracy	Precision	Recall	F1 Measure
VGG19	98.31%	0.982	0.983	0.983
VGG16	98.63%	0.984	0.984	0.981
DenseNet	95.32%	0.953	0.952	0.953
AlexNet	98.27%	0.976	0.977	0.977
GoogleNet	94.50%	0.942	0.943	0.942
LeNet	85.48%	0.849	0.853	0.851

Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, and Figure 10 show a comparison of the training and testing validation accuracies for VGG19, VGG16, DenseNet, AlexNet, GoogleNet, and LeNet, respectively. The comparison shows that both training and testing validation accuracies approach 100% as expected. These results indicate that the overfitting and the under-fitting problems were accounted for in this research with no under-fitting or overfitting problems reported. This is further proven and shown in Figure 11, Figure 12, Figure 13, Figure 14, Figure 15, and Figure 16 which show the comparison of the training and testing validation losses for VGG19. VGG16. DenseNet, AlexNet, GoogleNet, and LeNet, respectively. The comparison shows that both training and testing validation losses approached zero as expected.

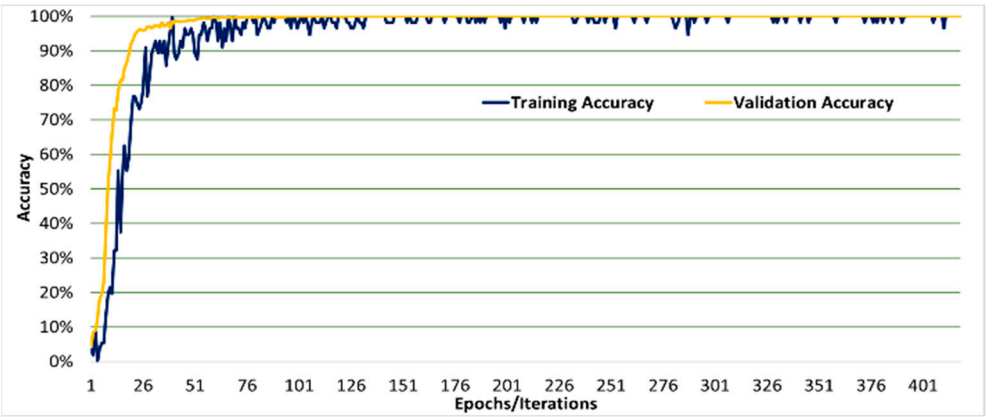


Figure 5. VGG19 accuracy learning curves for training and validation.

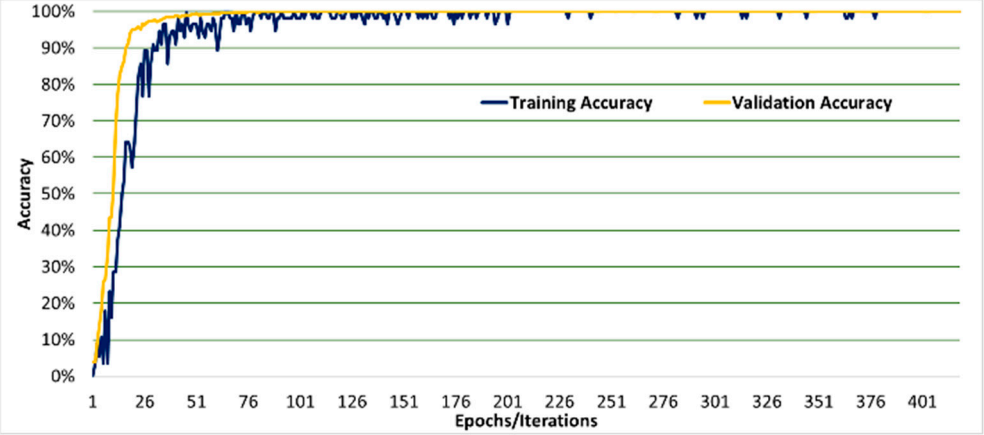


Figure 6. VGG16 accuracy learning curves for training and validation.

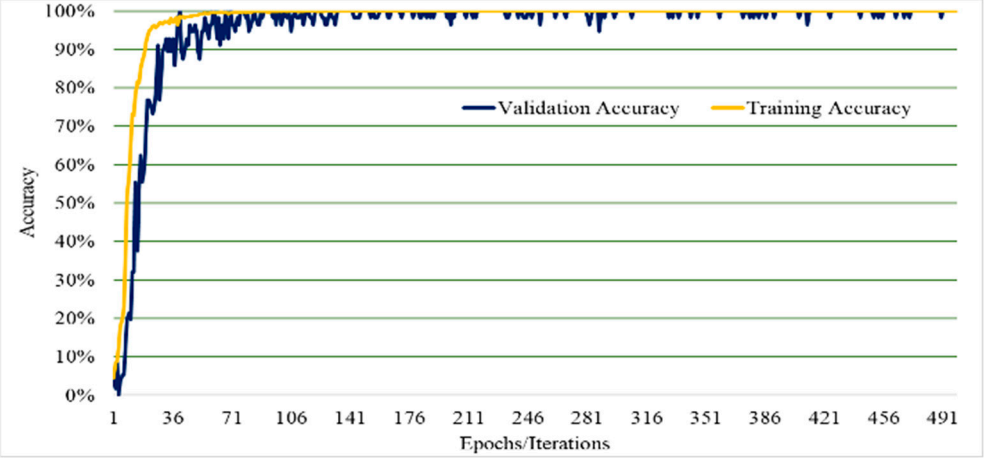


Figure 7. DenseNet accuracy learning curves for training and validation.

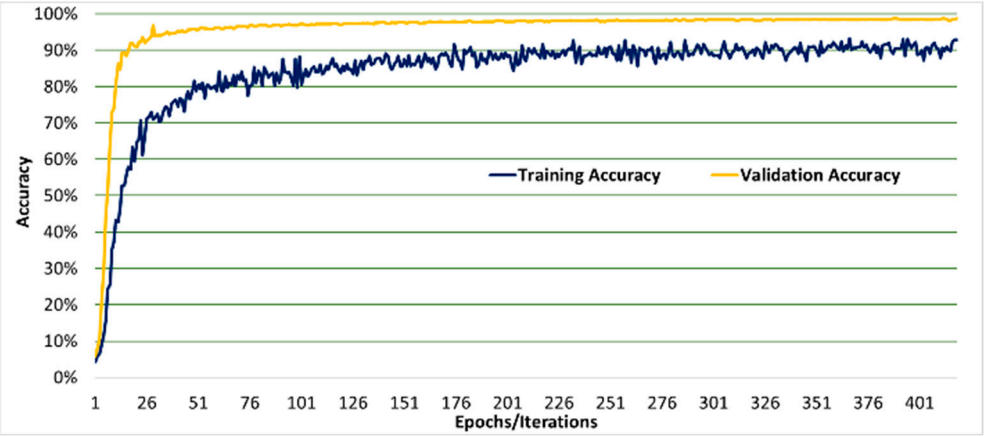


Figure 8. AlexNet accuracy learning curves for training and validation.

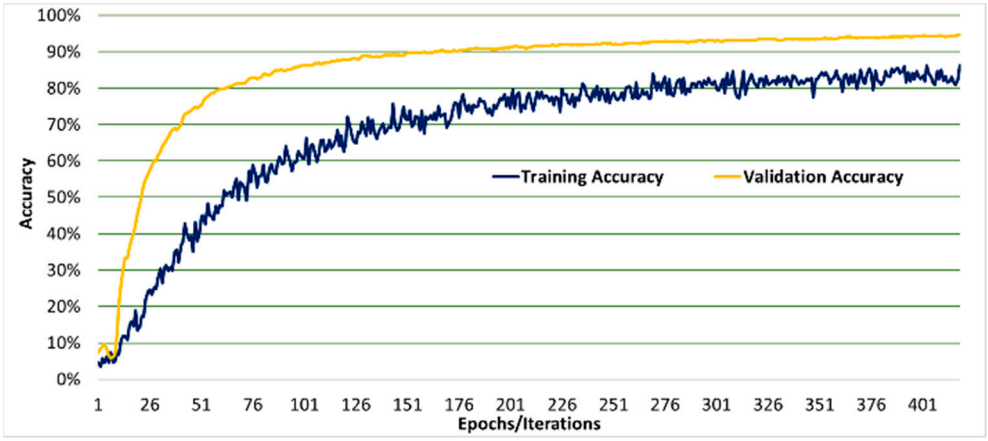


Figure 9. GoogleNet accuracy learning curves for training and validation.

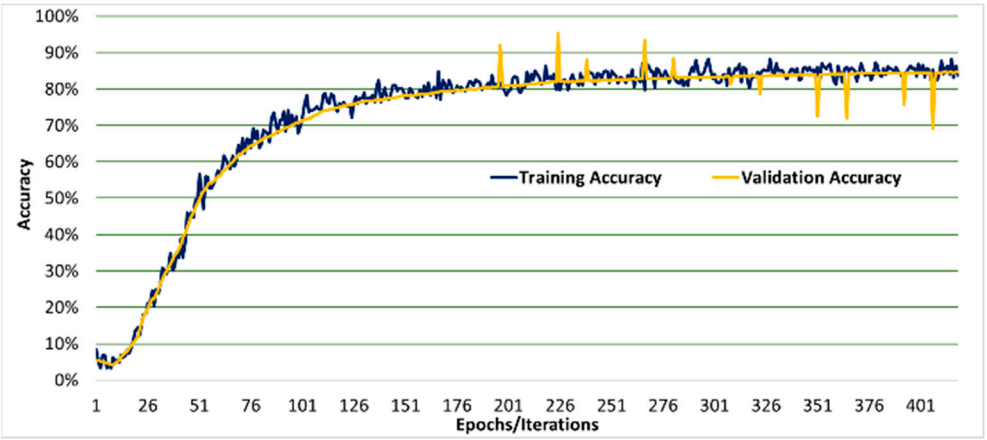


Figure 10. LeNet accuracy learning curves for training and validation.

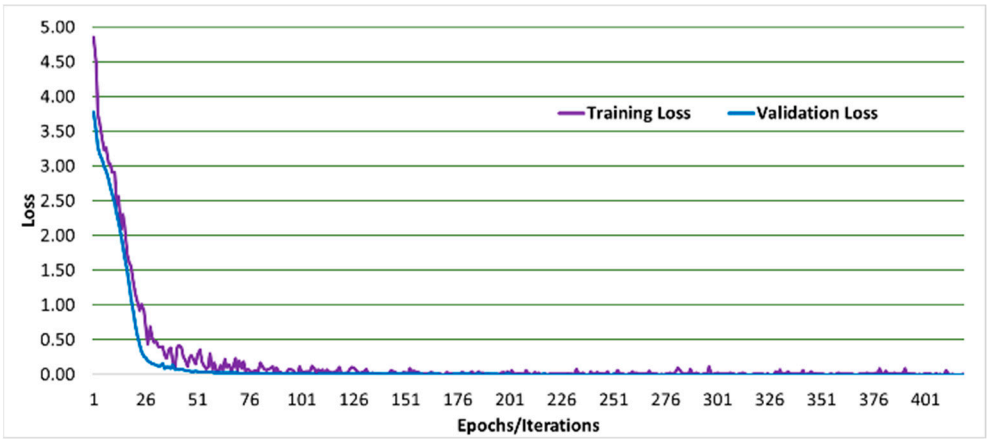


Figure 11. VGG19 loss learning curves for training and validation.

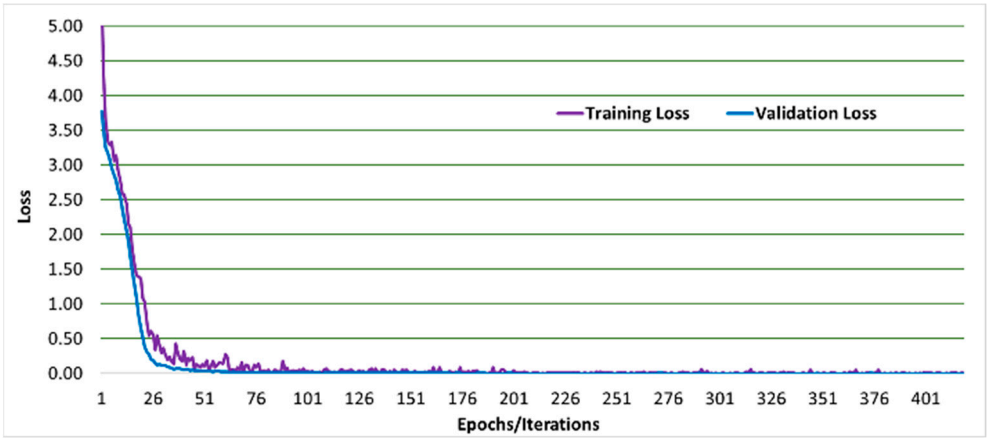


Figure 12. VGG16 loss learning curves for training and validation.

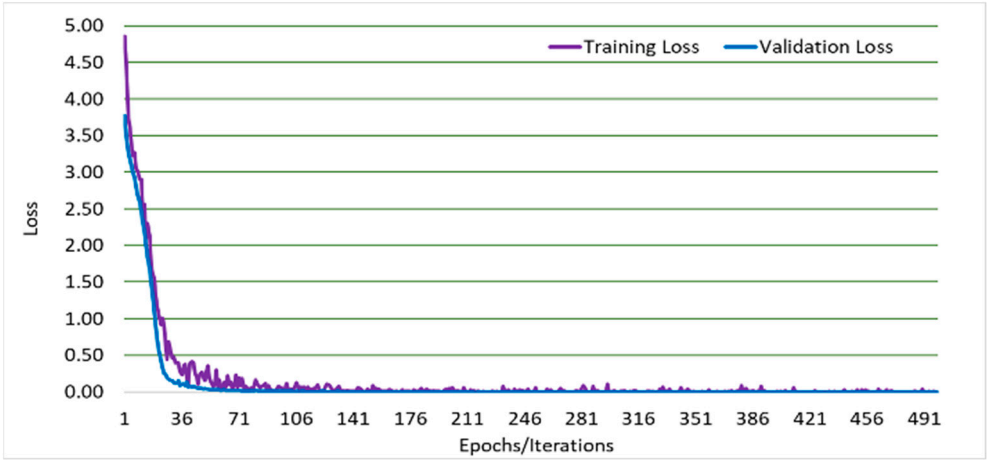


Figure 13. DenseNet loss learning curves for training and validation.

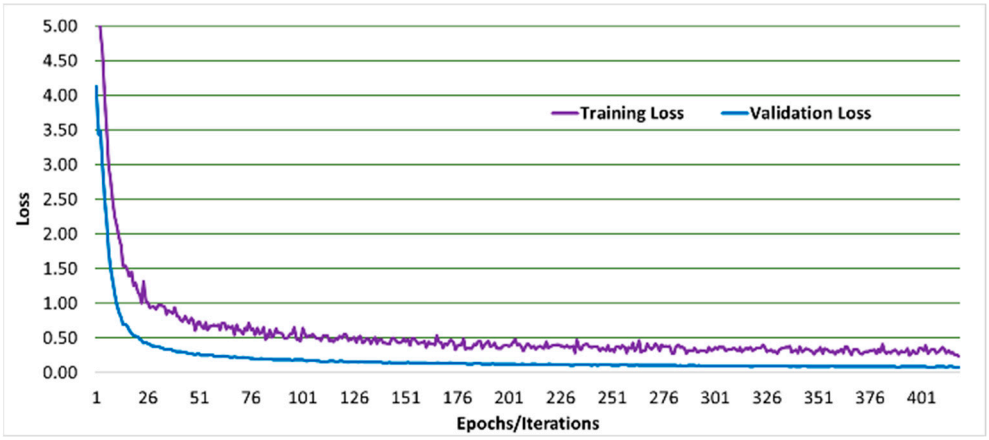


Figure 14. AlexNet loss learning curves for training and validation.

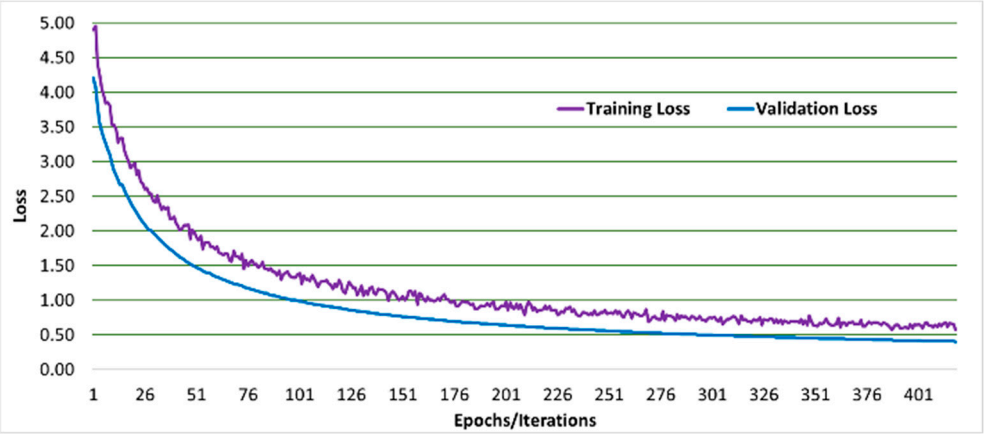


Figure 15. GoogleNet loss learning curves for training and validation.

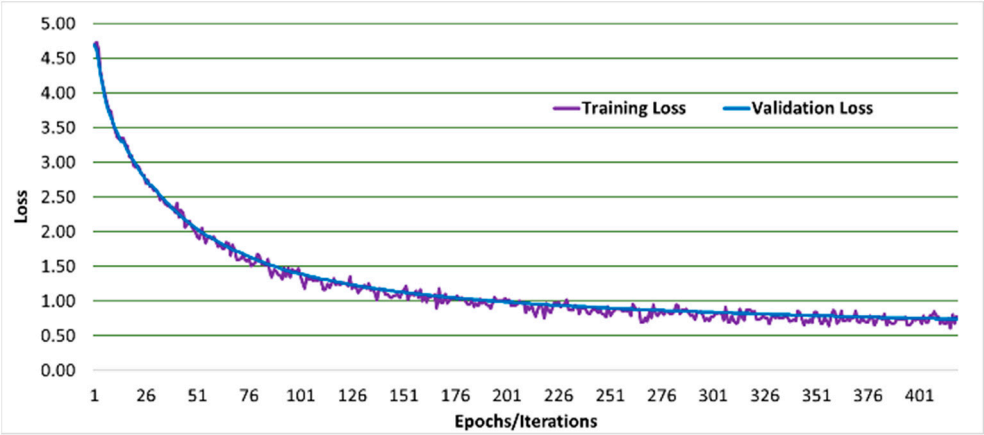


Figure 16. LeNet loss learning curves for training and validation.

The experiments were then repeated on the same optimized deep learning algorithms using the proposed non-freeze weights approach with the Arabic braille language dataset, as shown in Table 2. The accuracy increased significantly, with the best accuracy achieved by the VGG16 with an average accuracy of 99.68%, a precision of 98.36%, a recall of 97.96%, and an F1-measure of 98.16%. The lowest accuracy was still reported for GoogleNet with an average accuracy of 98.70%. Note that with non-freeze weights, the accuracy increased by 6.55% compared with the highest reported accuracy in Table 1. It should be clear here that the experiment was performed on the Arabic braille language dataset without augmentation.

Table 2. Experimental results of the proposed non-freeze weight-based CNN models for the Arabic braille language data.

	Accuracy	Precision	Recall	F1 Measure
VGG19	99.28%	0.991	0.992	0.992
VGG16	99.68%	0.993	0.991	0.994
DenseNet	96.51%	0.960	0.964	0.965
AlexNet	99.13%	0.989	0.988	0.989
GoogleNet	98.70%	0.984	0.981	0.981
LeNet	86.23%	0.858	0.860	0.861

The experiment is then repeated using the optimized deep learning algorithms using the proposed non-freeze weight approach but this time on the combined Arabic braille language dataset with the augmented dataset. The results are shown in Table 3. The results

indicate yet another increase in accuracy due to expanding the dataset size. The increase of 0.3% is actually significant as compared to results in Table 2 and dramatically significant as compared to results in Table 1 where the difference is 1.35%. The increase in accuracy is extremely important because this a proposed system that will serve for assistive learning for the visually impaired and they have no way of comparing the audio translation with the original unless they go to the traditional time-consuming touch-and-feel approach. Table 3 shows that the highest reported accuracy was again achieved using VGG16 with an average accuracy of 99.98%, precision of 99.4%, recall of 99.5%, and F1-measure of 99.7%. The lowest accuracy is again reported using th GoogleNet with an average value of 88.5%.

Table 3. Experimental results of the proposed non-freeze weight-based CNN models for the Arabic braille language augmented data.

	Accuracy	Precision	Recall	F1 Measure
VGG19	99.58%	0.995	0.994	0.995
VGG16	99.98%	0.994	0.995	0.997
DenseNet	98.62%	0.981	0.982	0.980
AlexNet	99.45%	0.993	0.991	0.989
GoogleNet	88.50%	0.883	0.879	0.881
LeNet	85.51%	0.850	0.851	0.851

The confusion matrix-based comparison obtained for the various experiments performed above with the best-performing VGG16 model is shown in Figures 17–19. These are the confusion matrices for the experiments performed on the Arabic braille language dataset. Figure 17 shows the confusion matrix for the basic VGG16 applied to the Arabic braille language dataset. It is noticed that even with the basic VGG16, the accuracy is high but it can be optimized to achieve better results because the application we are targeting is for the specific purpose of assistive learning technology for the visually impaired. Thus, an optimal solution can only be achieved as we approach approximately 100% on various complex datasets of the Arabic braille language dataset. Figure 18 shows the confusion matrix using the Optimized VGG16 model with the proposed transfer learning approach. It is noticed that the confusion matrix showed better results but still can stand for improvement for the optimal solution. Therefore, Figure 19 shows the confusion matrix using the optimized VGG16 along with the proposed transfer learning approach, which resulted in a further increase of accuracy.

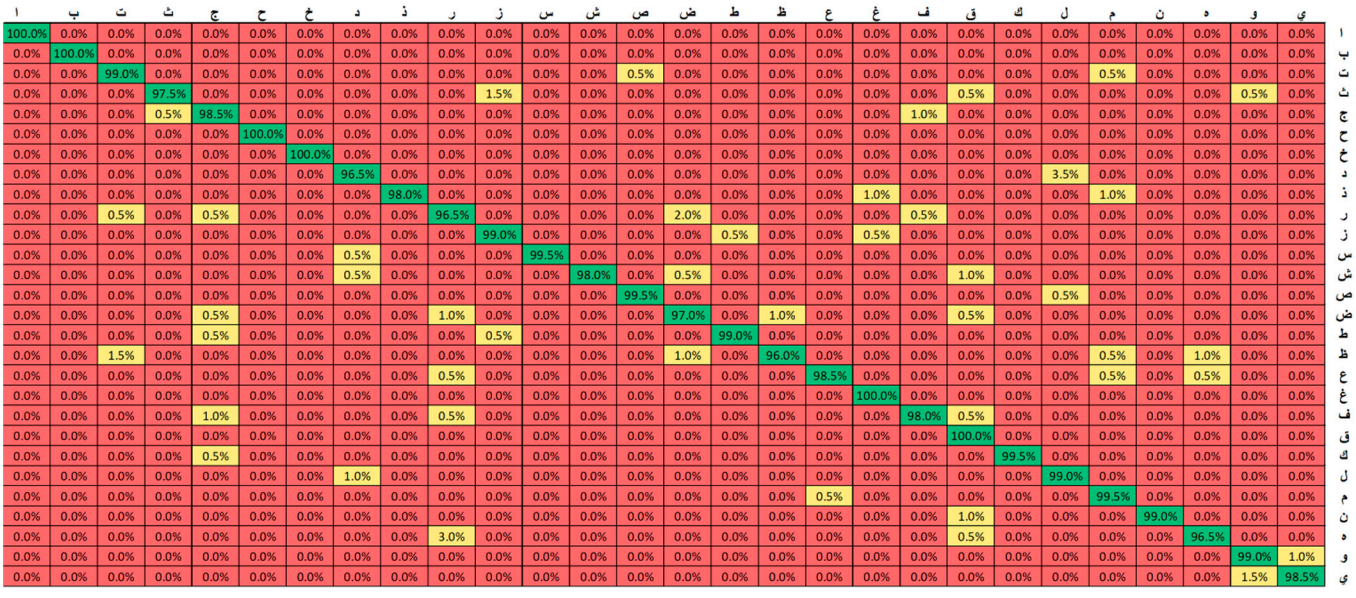


Figure 17. Confusion matrix for 28 Arabic braille letters classification using VGG16 basic transfer learning.

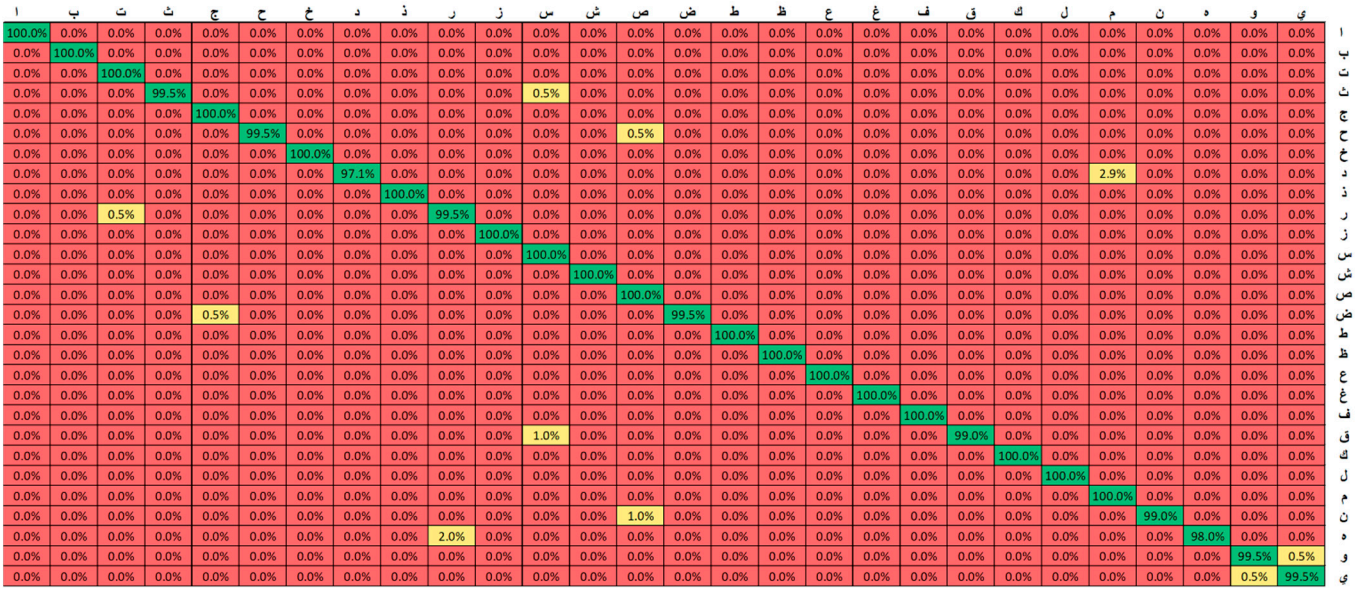


Figure 18. Confusion matrix for 28 Arabic braille letters classification using VGG16 proposed transfer learning.

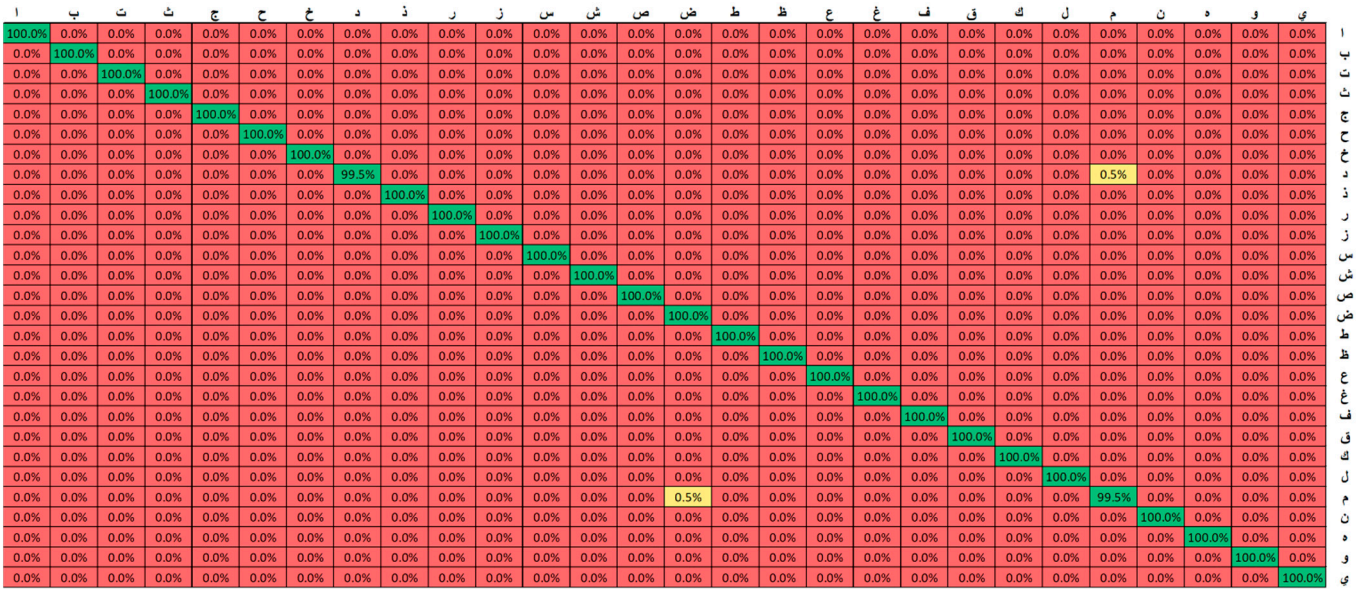


Figure 19. Confusion matrix for 28 Arabic braille letters classification using VGG16 proposed transfer learning with data augmentation.

Similarly, the best-performing model has also been tested using the English braille language dataset. According to Table 4, the highest achieved accuracy was 99.92% by Vgg16. Note that Table 4 shows the results of the experiment of applying the proposed non-freeze weight approach with the optimized CNN models on the combined dataset of the English braille language dataset with augmentation. The highest accuracy was achieved using VGG16 and reported as 99.92%, with a precision of 99.5%, recall of 99.4%, and Fe-Score of 99.5%. The lowest accuracy of 86.79% was reported when using the LeNet. The VGG16 took 5 h and 20 min for training which is slightly less than the VGG19 model and relatively more than other compared models. It has been noticed that the individual braille image test computational time was approximately the same for all models on the proposed device.

Table 4. Experimental results of the proposed non-freeze weight-based CNN models for the English braille language 26 letters augmented data.

	Accuracy	Precision	Recall	F1 Measure
VGG19	99.60%	0.991	0.994	0.994
VGG16	99.92%	0.995	0.994	0.995
DenseNet	97.46%	0.969	0.971	0.968
AlexNet	98.37%	0.970	0.970	0.97
GoogleNet	98.21%	0.976	0.978	0.977
LeNet	86.79%	0.864	0.863	0.863

5. Conclusions

Individuals with disabilities should continue to receive the utmost support since with the right education and tools they have proved themselves to be valuable members of the community. They contributed in many fields and history has recorded many famous individuals with disabilities and persons with visual impairment. They became famous because they achieved things that persons without disabilities and persons with full eyesight have not been able to achieve. Therefore, society must continue to support persons with disabilities to achieve their full potential. With the advancements in technology, many devices can be developed to assist persons with disabilities in the education field to enhance their education, learning, and knowledge. These technology-enhanced devices can assist them to learn or speed up their learning process. In this paper, we proposed an AI-based device that can automatically translate Arabic and English braille to the corresponding audio. This device can serve either Arabic-speaking individuals, English-speaking individuals, or bilingual individuals. There are many benefits to this device, including but not limited to teaching visually impaired individuals the braille language, teaching the relatives of the visually impaired individual the braille language, or assisting visually impaired individuals who already know braille to read braille documents/books at much faster speeds. The proposed system optimized deep learning models along with transfer learning. The main idea of the proposed system is to optimize the deep learning algorithms and then freeze the first portion of layers and unfreeze the second portion of the layers to allow the systems to retrain and update the weights accordingly. This resulted in an enhanced accuracy for both the Arabic language braille and English language braille. In addition, increasing dataset size and complexity allows for better performance. Therefore, augmentation was performed for both the Arabic and English braille language datasets to increase the dataset sizes. The increased sizes of datasets using the proposed method resulted in even higher optimal accuracies.

Future work in this field will include field testing the device to receive actual feedback from individuals with visual impairments. The system will continue to be enhanced based on the feedback from individuals with visual impairments and their relatives.

Author Contributions: G.L. conducted this research; G.B.B. worked on the methodology and analysis; R.A., S.E.A. and G.A. did the initial writing; R.A. reviewed the paper and fixed grammar; R.A. and K.A. helped with the literature review. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Prince Mohammad bin Fahd Futuristic Studies Research Grant 2022. This work was also supported in part by the Commonwealth Cyber Initiative, an investment in the advancement of cyber R&D, innovation, and workforce development. For more information about CCI, visit https://cyberinitiative.org.

Data Availability Statement: The data used in this research was newly created which can be acquired on request by sending email to glatif@pmu.edu.sa.

Acknowledgments: All authors acknowledge the support from the Prince Mohammad Bin Fahd University for providing the computational resources to conduct this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shokat, S.; Riaz, R.; Rizvi, S.S.; Abbasi, A.M.; Abbasi, A.A.; Kwon, S.J. Deep Learning Scheme for Character Prediction with Position-Free Touch Screen-Based Braille Input Method. *Hum.-Cent. Comput. Inf. Sci.* **2020**, *10*, 41. [CrossRef]
2. Latif, G.; Mohammad, N.; AlKhalaf, R.; AlKhalaf, R.; Alghazo, J.; Khan, M. An Automatic Arabic Sign Language Recognition System Based on Deep CNN: An Assistive System for the Deaf and Hard of Hearing. *Int. J. Comput. Digit. Syst.* **2020**, *9*, 715–724. [CrossRef]
3. Khan, S.; Rahmani, H.; Shah, A.; Bennamoun, M. *A Guide to Convolutional Neural Networks for Computer Vision*; SpringerLink: Berlin/Heidelberg, Germany, 2018.
4. Alufaisan, S.; Albur, W.; Alsedrah, S.; Latif, G. Arabic Braille Numeral Recognition Using Convolutional Neural Networks. *Springer eBooks* **2021**, *9*, 87–101.
5. Tiendee, S.; Lerdsudwichai, C.; Thainimit, S.; Sinthanayothin, C. The Method of Braille Embossed Dots Segmentation for Braille Document Images Produced on Reusable Paper. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 163–170. [CrossRef]
6. Shokat, S.; Riaz, R.; Rizvi, S.S.; Khan, I.; Paul, A. Detection of Touchscreen-Based Urdu Braille Characters Using Machine Learning Techniques. *Mob. Inf. Syst.* **2021**, *2021*, 1–16. [CrossRef]
7. Shokat, S.; Riaz, R.; Rizvi, S.S.; Khan, I.; Paul, A. Characterization of English Braille Patterns Using Automated Tools and RICA Based Feature Extraction Methods. *Sensors* **2022**, *22*, 1836. [CrossRef]
8. Perera, T.D.S.H.; Wanniarachchi, W.K.I.L. Optical Braille Recognition Based on Histogram of Oriented Gradient Features and Support-Vector Machine. *Int. J. Eng. Sci. Comput.* **2018**, *8*, 19192–19195.
9. Asebriy, Z.; Raghay, S.; Bencharef, O. An Assistive Technology for Braille Users to Support Mathematical Learning: A Semantic Retrieval System. *Symmetry* **2018**, *10*, 547. [CrossRef]
10. Sufiun, A.; Jabiullah, M.I. A Novel Approach of CNN Patterns Extraction for Bangla Handwriting to Bangla Braille Notation. *Int. J. Eng. Adv. Res.* **2021**, *3*, 1–15.
11. Jha, V.; Parvathi, K. Braille Transliteration of hindi handwritten texts using machine learning for character recognition. *Int. J. Sci. Technol. Res.* **2019**, *8*, 1188–1193.
12. Prakash, S.; Thomas, S.; Gopalan, S.M. An Effective Approach of English Braille to Text Conversion for Visually Impaired Using Machine Learning Technique. *EasyChair Prepr.* **2023**, *9908*, 1–9.
13. Souza, M.D.; Preetham, S.; Varun, S.M.; Vardhan, N.; Venkatraman, G. Braille Character Recognition Using Deep Learning Strategy Image Processing and Computer Vision. *Int. Res. J. Mod. Eng. Technol. Sci.* **2023**, *5*, 6385–6389.
14. Chellaswamy, C.; Geetha, T.S.; Hariharan, K.; Archana, K.; Babitharani, S. Deep Learning-Based Braille Technology for Visual and Hearing Impaired People. In Proceedings of the 2023 International Conference on Smart Systems for Applications in Electrical Sciences, Tumakuru, India, 7–8 July 2023; pp. 1–8.
15. Fogarassy-Neszly, P.; Pribeanu, C. Multilingual text-to-speech software component for dynamic language identification and voice switching. *Stud. Inform. Control.* **2016**, *25*, 335–342. [CrossRef]
16. Bhatia, S.; Devi, A.; Alsuwailem, R.I.; Mashat, A. Convolutional Neural Network Based Real Time Arabic Speech Recognition to Arabic Braille for Hearing and Visually Impaired. *Front. Public Health* **2022**, *10*, 898355. [CrossRef] [PubMed]
17. Latif, G.; Alghmgham, D.A.; Maheswar, R.; Alghazo, J.; Sibai, F.; Aly, M.H. Deep Learning in Transportation: Optimized Driven Deep Residual Networks for Arabic Traffic Sign Recognition. *Alex. Eng. J.* **2023**, *80*, 134–143. [CrossRef]
18. Mohammed, A.S.; Hasanaath, A.A.; Latif, G.; Bashar, A. Knee Osteoarthritis Detection and Severity Classification Using Residual Neural Networks on Preprocessed X-ray Images. *Diagnostics* **2023**, *13*, 1380. [CrossRef]
19. Saleem, M.A.; Senan, N.; Wahid, F.; Aamir, M.; Samad, A.; Khan, M. Comparative Analysis of Recent Architecture of Convolutional Neural Network. *Math. Probl. Eng.* **2022**, *2022*, 1–9. [CrossRef]
20. Tao, Y.; Xu, M.; Lu, Z.; Zhong, Y. DenseNet-Based Depth-Width Double Reinforced Deep Learning Neural Network for High-Resolution Remote Sensing Image Per-Pixel Classification. *Remote Sens.* **2018**, *10*, 779. [CrossRef]
21. Latif, G.; Morsy, H.A.; Hassan, A.; Alghazo, J. Novel Coronavirus and Common Pneumonia Detection from CT Scans Using Deep Learning-Based Extracted Features. *Viruses* **2022**, *14*, 1667. [CrossRef]
22. Qu, Y.; Tang, W.; Feng, B. Paper Defects Classification Based on VGG16 and Transfer Learning. *J. Korea TAPPI* **2021**, *53*, 5–14. [CrossRef]

Computer graphics for the blind

Satoshi INA

Tsukuba College of Technology
4-12 Kasuga,Tsukuba,305,JAPAN

Abstract

This paper presents a tactile hard-copy system which help an "active tactile graphics" for the blind. Here the word "active" means that the blind can make graphics programs and deal with computer graphics by his own efforts. It translates the color graphics on the screen into an embossed dots image on paper. To get a precise embossed dots hard-copy close to the real graphics screen, we adopted a Braille printer plotter. The program stays resident in computer's memory, and a push of COPY key at any time starts to make an embossed hard-copy of the current graphic screen. By this system and computer language like C, the blind can try to make graphics programs and examine the colors and the graphic figures constructing the screen graphics. The system can extract several combinations of colors from the whole graphics selectively and make the hard-copys. This color selective hard-copys help the blind understand the whole graphics as a result, because each hard-copy reduces the complexity of the graphics and tends to be easier to examine by touch.

1 Introduction

We have developed a TSR(Terminate and Stay Resident) hard-copy system that translates color pixels on the screen into embossed dots on paper by Braille printer plotter. Recently color graphics capacity in Personal Computer(PC) and programming language like C is more and more powerful. Even under the DOS environment, many programs handle not only characters but also color graphics or images. The blind people can recognize almost all the character information on the screen through on-line screen reader using voice synthesizer and Braille display. (Braille display is an on-line translator from character to Braille.) On the other hand there are not good methods to recognize graphics or images on the screen. Especially real time methods are not put in practical use. We know several methods to present graphics for the blind, such as the screen search by OPTACON[4] or relief picture on micro capsule paper. We have already referred to the problems they have, and presented the system(tenzu,brlview) which give the sighted persons an easy way to make tactile graphics on paper for the blind[1][2]. We call this "passive tactile graphics" because the blind cannot make graphics by himself. And now, instead of the passive tactile graphics, we are going to attack "active tactile graphics". It means that the blind can recognize the graphics on the screen by himself and they can make color graphics program by himself.

2 Purpose

Our purpose is to make a system that support active tactile graphics. From this point of view we have made a graphics screen hard-copy system onto a Braille printer plotter. We call this system BHCOPY hereafter. BHCOPY makes the following matters possible through the embossed dots hard-copy.
(1)The blind can feel graphics and colors on the PC's screen
(2)The blind can make color graphics program on the PC's screen by himself

Picture 1. ESA721 Braille printer plotter

3 System Configuration

(1)PC(NEC PC9801FA) OS:MS-DOS Ver.3.3C,CPU:Intel80486SX,CLOCK:16MHz
(2)Braille printer plotter(NEW ESA721)[3]. Picture 1 shows the external appearance.The Braille printer plotter has 3 sizes of dot(Small/Medium/Large). The resolution is referred in section 5.

4 Method

Hard-copy originally means to copy all characters and graphics on screen onto paper as they are in the same resolution. Our personal computer(NEC PC9801series) has two types of memories as display memory, one is graphic VRAM(GVRAM) and the other is text VRAM(TVRAM). TVRAM reserves character codes to display characters on the screen. Some screen readers peek this TVRAM to read the screen and speak. Our hard-copy system BHCOPY peeks GVRAM. GVRAM reserves graphics pixel information on the screen. It has three color planes corresponding to three primary colors(Red,Green,Blue). So the number of colors is 8. The resolution of the PC's screen is 640*400 pixels. On the other hand, Braille printer plotter's resolution is 792*599(step size:0.34mm) but the effective resolution is far lower than that of the screen by the reason mentioned in the following section 5.
To compensate for the lack of resolution, the next operations are effective in the hard-copy process.
(1)hard-copy of the discriminated colors on the screen
(2)hard-copy of the separated colors on the screen
(3)hard-copy by superposition of the selected colors on the screen
(4)selection and control of dot-size and dot-interval on the Braille printer plotter

5 Resolution

The difference of effective resolutions between graphic screen and Braille printer plotter is very large as shown below.

(1)The resolution of graphics screen

Horizontal:<640

Vertical :<400

(2)The resolution of the Braille printer plotter

Plotter step size is about 0.34mm, so the resolution is as follows.

793*600 for 10*11 inch Braille paper

726*480 for 8*10 inch Braille paper

The resolution seems enough to plot embossed dot graphics. But the dots have to be embossed suitably isolated each other not to break paper. It means we have to put intervals of 3 or 5 steps between the dots. So the effective resolution which do not break paper is regarded as the value shown in the following clause (3).

(3)The effective resolution of Braille image for 8*10 inch Braille paper

Horizontal:145-242

vertical :96-160

6 Functions and how to use

We have two types of mode to make a screen hard-copy. One is a TSR(Terminate and Stay Resident) executable program, and the other is a subroutine callable objective program.

(1)TSR program mode(named BHCOPY)

To load and stay resident BHCOPY program, bhcopy command needs to be executed. Once the program is resident, a push of the COPY key at any time starts to make an embossed dots hard-copy of the graphics screen. BHCOPY has several options to change the hard-copy mode. The important options are color selection, dot-s size selection, and dot-interval selection. To change those parameter options, once you haft release the residence and execute the command with new parameters again. The command format is as follows.

(a)Command format

bhcopy [-r] [-?] [-d{0|1|2}] [-i{3|5|7|9}] [-c{a|r|g|b|y|c|m|w|f}]

Default parameters of bhcopy are regarded as "bhcopy -d0 -i5 -ca".

Meaning of the option parameters is as follows.

r-remove the resident program

?-show how to use

d-select dot size(0:Small,1:Medium,2:Large)

i-select dot interval(1unit(step)=0.34mm)

c-select color to make hard-copy

'a':All color,'r':Red,'g':Green,'b':Blue,

'y':Yellow,'c':Cyan,'m':Magenta,'w':White,

'f':read color-dot one-to-one correspondence table from file(col2dot.tab)

(b)Command sample

A>bhcopy -d0 -i3 -cy<return>

Only yellow pixels are extracted and hard-copied using small size dot with interval 3.

A>bhcopy -r<return>

BHCOPY program is removed from memory.

(2)Subroutine callable mode(named GHCOPY)

In this mode users can link the GHCOPY object module and call the function in their program. The calling format is as follows. The meaning of parameters is same as the TSR program mode.

(a)Calling format in C language

void ghcopy(int dotsize,int interval,char color)

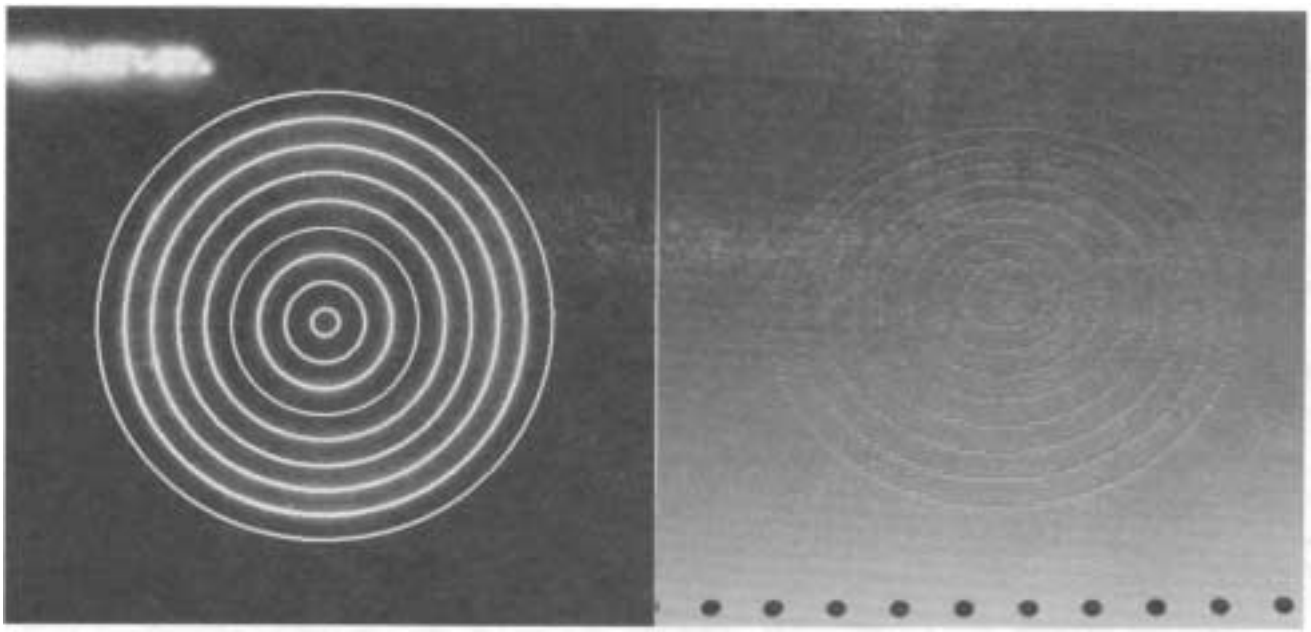
(b)Sample Program

```
void main(void) {ghcopy(0,5,'w');ghcopy(1,7,'g');}
```

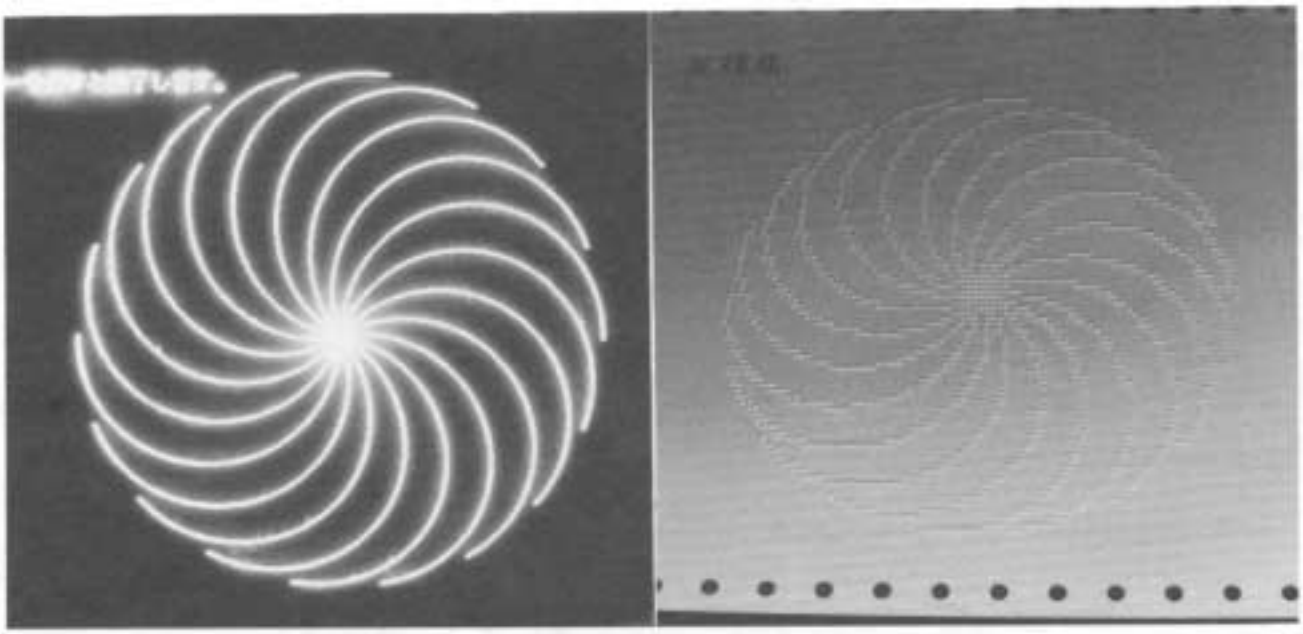
This code makes embossed graphics hard-copy by converting white pixels into small-size embossed dots, and green pixels into medium-size embossed dots.

7 Examples

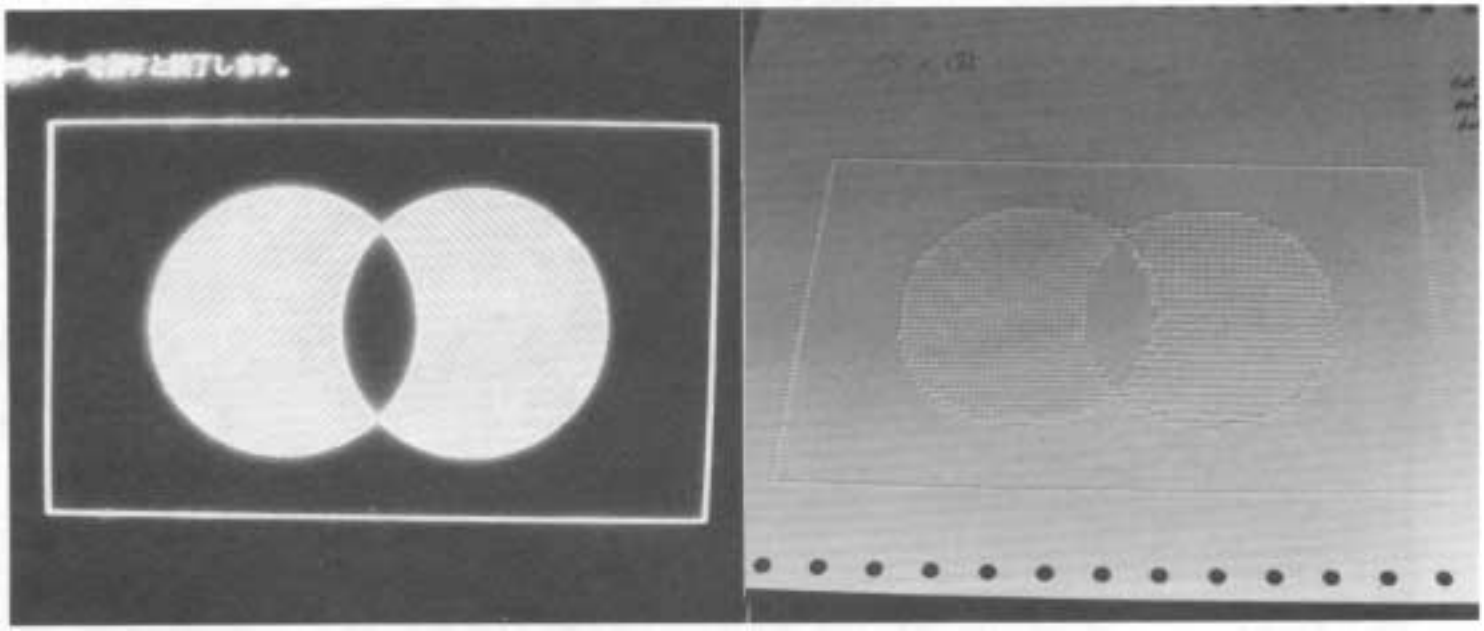
Some examples(named Flower,Concentric circles,Triangular and Quadratic wave,Venn diagram,sin/cos wave) are shown in Picture 2 to 6. They are shown as the pair of an original graphics on the screen and an embossed dots hard-copy. Only hard-copy outputs of a pie chart and a bar chart are shown in Picture 7,8. The original graphics are color and sampled and translated into an embossed dots pattern including one or three types of dot size by selection of option parameter.



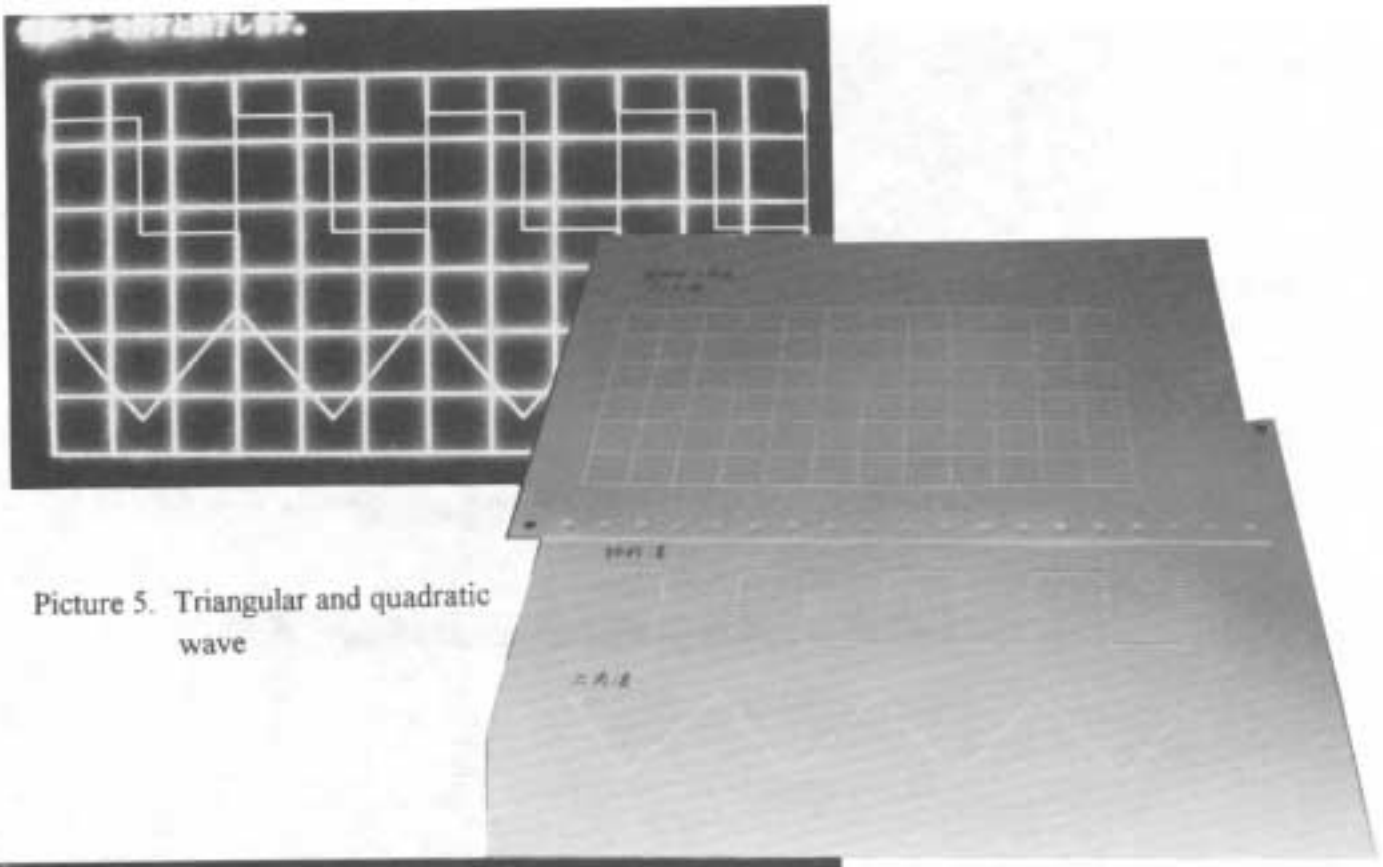
Picture 2. Concentric circles



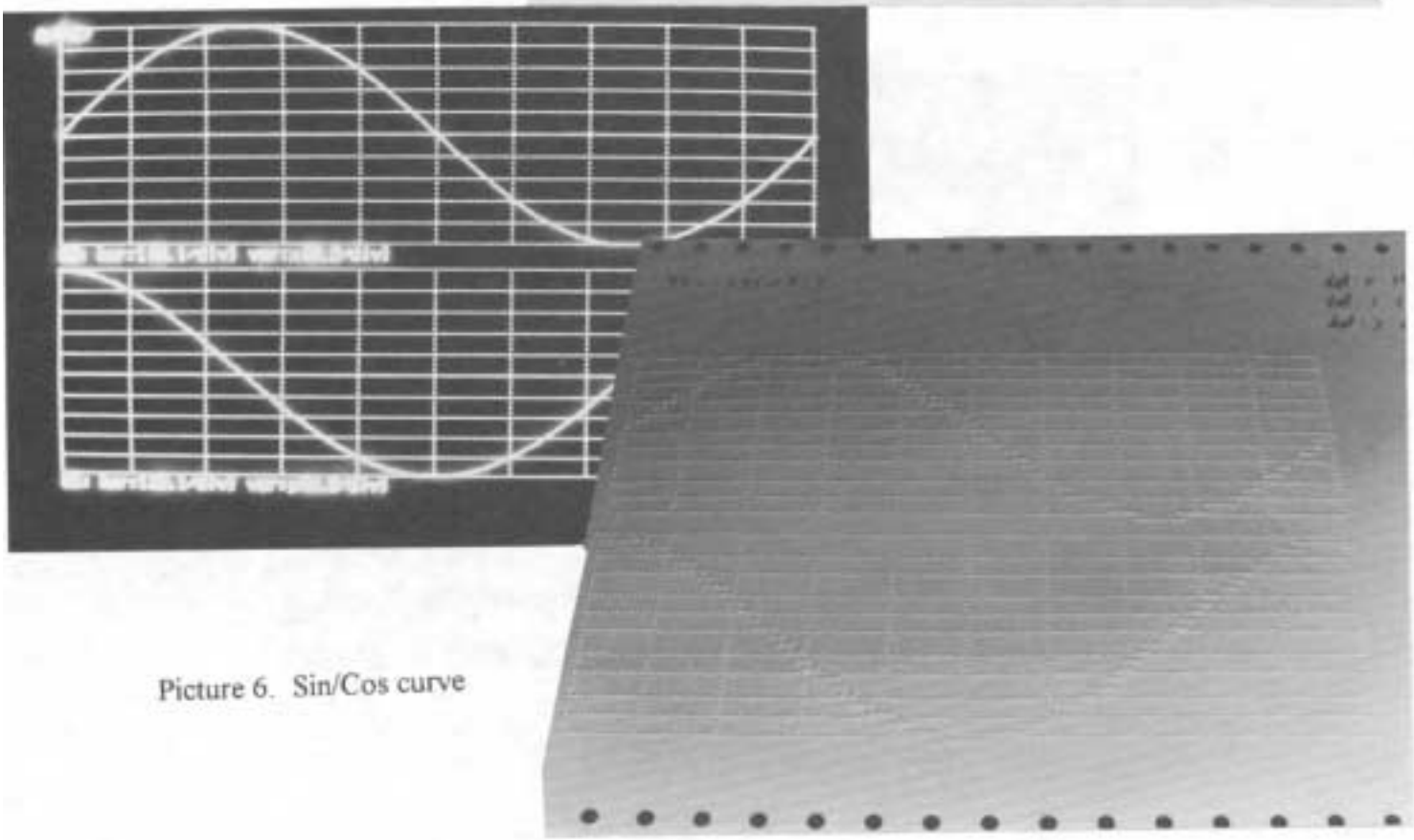
Picture 3. Flower



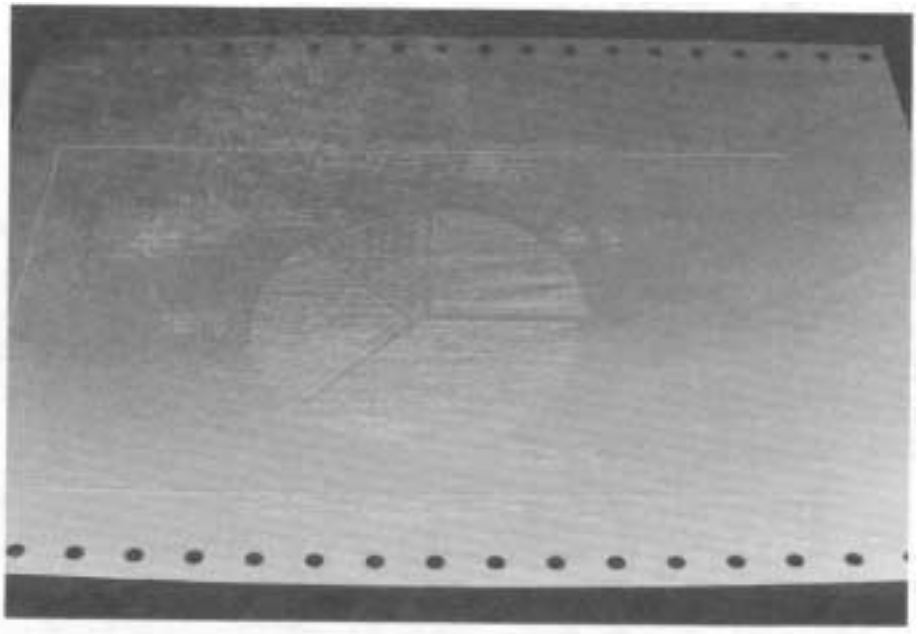
Picture 4. Venn diagram



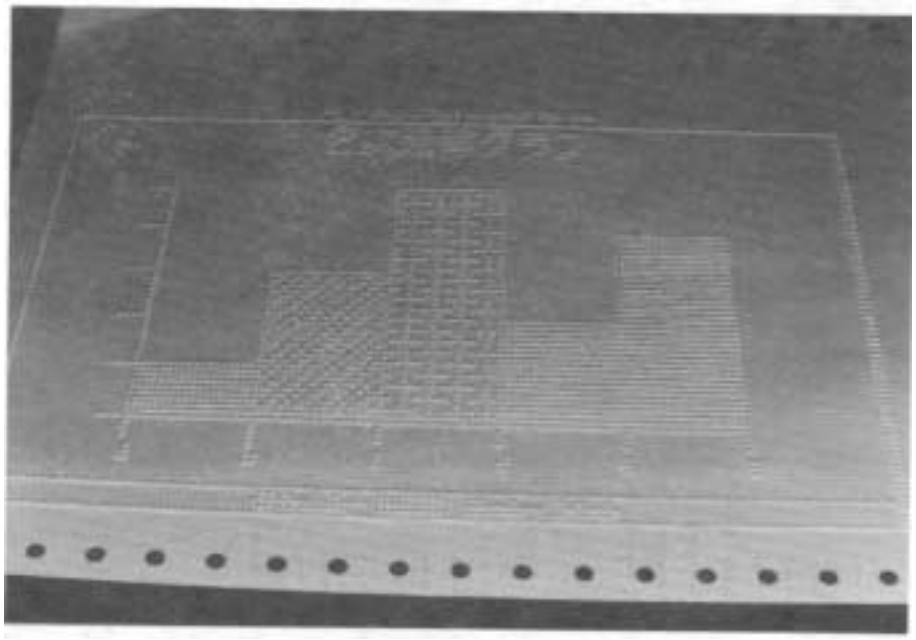
Picture 5. Triangular and quadratic wave



Picture 6. Sin/Cos curve



Picture 7. A hard-copy of pie chart



Picture 8. A hard-copy of bar chart

8 Result and discussion

We have proposed a prototype system for the "active tactile graphics for the blind" as an embossed dots hard-copy system on PC. And we find out the following facts and make them possible in our system.

Color graphics on the screen tend to be colorful and complex for the blind to recognize at once by tactile sensation. So that the blind can recognize these complex graphics by tactile sensation, simplification process of the picture elements is essential. From this point of view we propose to decompose the graphics on the screen into several elements according to colors and produce each color pattern as a hard-copy. This process not only simplifies the color graphics but also help the blind find out the color organization of the graphics. After that decomposition and recognition process, the superposed hard-copy of all color elements can also be understood as a whole graphics.

We are often asked by blind persons to translate graphics characters into Braille characters in a hard-copy process. But I have to say it's too hard. The reason is as follows. On the GVRAM, there are no character and geometric information left such as character codes, dot, line, rectangle, circle and so on other than pixels. To make this possible, maybe a kind of high performance pattern recognition method has to be put in.

References

[1]Satoshi INA:Development of 2D Tactile Graphics Editor and Printing System for Document with Braille and Graphics.,IEICE EIC,Vol.J77-D-II,No.10,pp1973-1983,1994.(in Japanese)

[2]Satoshi INA:Personal computer aided tactile graphics system.,International Conference on higher education for students with disabilities,ABSTRACTS pp.C-II-c,1993.

[3]JTR Inc.:NEW ESA721braille printer user's manual,1990

[4]Telesensory Systems Inc.,OPTACON user's manual

CONVERSION OF BRAILLE TO TEXT IN ENGLISH, HINDI AND TAMIL LANGUAGES

S.Padmavathi[1], Manojna K.S.S[2], Sphoorthy Reddy .S[3] and Meenakshy.D[4]

Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

[1]s_padmavathi@cb.amrita.edu, [2]manojna.kapala@gmail·com,
[3]sphoorthy.surakanti@gmail.com, [4]meenakshymenon14@gmail.com,

ABSTRACT

.The Braille system has been used by the visually impaired for reading and writing. Due to limited availability of the Braille text books an efficient usage of the books becomes a necessity. This paper proposes a method to convert a scanned Braille document to text which can be read out to many through the computer. The Braille documents are pre processed to enhance the dots and reduce the noise. The Braille cells are segmented and the dots from each cell is extracted and converted in to a number sequence. These are mapped to the appropriate alphabets of the language. The converted text is spoken out through a speech synthesizer. The paper also provides a mechanism to type the Braille characters through the number pad of the keyboard. The typed Braille character is mapped to the alphabet and spoken out. The Braille cell has a standard representation but the mapping differs for each language. In this paper mapping of English, Hindi and Tamil are considered.

KEYWORDS

Braille Conversion, Projection Profile, Tamil Braille conversion, Hindi Braille conversion, Image Segmentation

1. INTRODUCTION

Visually impaired people are an integral part of the society. However, their disabilities have made them to have less access to computers and Internet than the people with clear vision. Over time Braille system has been used by them for written communication. Braille is a system of writing that uses patterns of raised dots to inscribe characters on paper. This allows visually impaired people to read and write using touch instead of vision. It is the way for blind people to participate in a literate culture. First developed in the nineteenth century, Braille has become the pre-eminent tactile alphabet. Its characters are six- dot cells, two wide by three tall as shown in figure 1. Each dot may exist or may not exist giving two possibilities for each dot cell. Any of the six dots may or may not be raised; giving 64 possible characters. In English it includes 26 English alphabets, punctuations, numbers etc. Figure 2 shows the Braille representation of English alphabets. Braille representation for numerals is shown in figure 3.

DOI : 10.5121/ijcsea.2013.3303

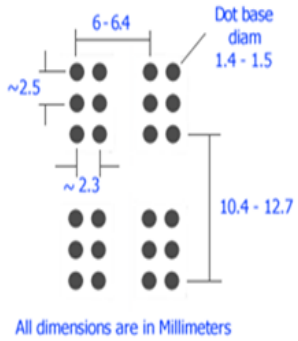


Fig1:Braille cell dimensions

Fig 2.Braille characters of Alphabet

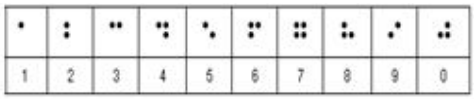


Fig 3. Braille characters for numerals

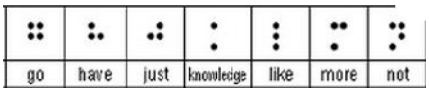


Fig 4. Contraction

Although Braille cells are used world-wide, the meaning of each cell depend on the language that they are being used to depict. In English Braille there are three levels of encoding: Grade 1, a letter-by-letter transcription used for basic literacy; Grade 2, an addition of abbreviations and contractions; and Grade 3, contains over 300 abbreviations and contractions that reduce the amount of Braille codes needed to represent written text. Some of the contracted words are represented in figure 4 in its Braille format.

Braille can be seen as the world's first binary encoding scheme for representing the characters of a writing system. However, very limited numbers of Braille books are available for usage. Printing of Braille books is a time consuming process. The requirement of special printers and software add to their limited availability. Scanned and text converted documents can be used in the meantime to serve the needs of the blind.

This paper mainly focuses on conversion of a Braille document into its corresponding alphabets of three main languages namely English, Tamil and Hindi using various concepts of image processing. The presence of dots in the Braille cells has to be identified to recognize the characters. The edge detection when applied on the scanned document will not produce the dots, hence the approximate intensity range of the dots are identified from the histogram. The image is treated through a sequence of enhancement steps which increases the contrast between the dots and the background. The edge detectors are then applied and the text area is cropped excluding

the borders through projection profile method. The document is then segmented into Braille cells using standard Braille measurements and projection profiles. The presence of dots in each cell is identified using a Threshold and converted to Binary sequence which is then mapped to the corresponding language alphabet. This paper also proposes a Number keypad which could be used for typing the analogous Braille alphabet using six numbers i.e. (7,4,1,8,5,2) corresponding to the six dot cells.

Section 2 of this paper discusses about few commercial systems available for converting Braille to text or vice-versa. Section 3 explains about the conversion of Braille to text; section 4 covers the outcomes of the proposed method. Section 5 concludes the paper.

2. LITERATURE SURVEY

A Braille translator is a software program that translates a script into Braille cells, and sends it to a Braille embosser, which produces a hard copy in Braille script of the original text. Basically only the script is transformed, not the language.

One of the general purpose translators is text to Braille converter. Few other commercial translators are also available such as win Braille, supernova, cipher Braille translator and Braille master.

2.1. Text to Braille converter:

Displays Braille as the user types characters. This convertor is OS independent and language used is java. It concentrates on conversion from English to Braille

2.2. Win Braille:

As referred in [2] Win Braille can be used without prior Braille knowledge. It includes standard Windows image control and the unique feature to convert images to tactile graphic format on-line.

2.3. Braille Master:

As referred in [5] the Braille Master package comes with both Windows and DOS versions. A large print facility suitable for partially sighted persons is also included in this package.

2.4. Cipher Braille Translator:

As referred in [2] Cipher is a text to Braille program that converts text documents into a format suitable for producing Braille documents, through the use of a Braille printer. The user can edit, save, use style templates and enable translation rules.

2.5. Supernova:

As referred in [2] Supernova is a window-based magnifier, screen reader and a Braille system that supports the conversion of text to speech, Braille displays and note-takers. Braille can be converted to text using number keypad and image processing techniques which are feasible for common people.[8] refers to a paper on Braille word segmentation and transformation of Mandarin Braille to Chinese character. [9] discuss the main concepts related to OBR systems;

list the work of different researchers with respect to the main areas of an OBR system, such as pre-processing, dot extraction, and classification. [10] describes an Arabic Braille bi-directional and bi-lingual translation/editor system that does not need expensive equipments. [11] focuses on developing a system to recognize an image of embossed Arabic Braille and then convert it to text. [12] presents a new Braille converter service that is a sample implementation of scalable service for preserving digital content. [13] proposes a software solution prototype to optically recognise single sided embossed Braille documents using a simple image processing algorithm and probabilistic neural network. [14] describes an approach to a prototype system that uses a commercially available flat-bed scanner to acquire a grey-scale image of a Braille document from which the characters are recognised and encoded for production or further processing. [15] introduces a new OBR system which designed for recognizing a scanned Arabic Braille document and converting it into a computerized textual form that could be utilized by converting it into voice using other applications, or it could be stored for later use. [16] presents an automatic system to recognize the Braille pages and convert the Braille documents into English/Chinese text for editing.[17] describes a new technique for recognizing Braille cells in Arabic single side Braille document. [18] describes the character recognition process from printed documents containing Hindi and Telugu text. [19] involves a keyboard which is a device made of logical switches and uses Braille system technique for sensing the characters[20] develope a system that converts, within acceptable consrtraints , (Braille image) to a computer readable form. [21] describes the Sparsha toolset. [22] presents a system for a design and implementation of Optical Arabic Braille Recognition(OBR) with voice and text conversion. [24] provides a detailed description of a method for converting Braille as it is stored as characters in a computer into print. [25] describes a new system that recognizes Braille characters from image taken by a high speed camera to Chinese character and at the same time automatically mark the Braille paper.

3. BRAILLE CONVERSION

For converting the Braille document to text, the input is taken in two different formats. In the first method the Braille character is accepted as a sequence of numbers typed through the keypad and in the second method a scanned Braille document is taken as input. The Braille character is extracted in each case and matched with the corresponding alphabet with help of a pre built Trie structure.

3.1. Keypad to type Braille document

The six dot cell representation of Braille character could be numbered from 1 to 6 starting from top left to bottom right in the order left to right and top to bottom. The numbers 7,4,1,8,5,2 of keypad are mapped to the dots 1,2,3,4,5,6 respectively as shown in fig 5.

Fig 5. Mapping of dots to numbers

With the number pad the number sequences of the Braille characters are typed and used for further conversion. The number sequences of the English Braille alphabets are listed in Table 1 and that of contractions are shown in Table 2.

Table 1: Mapping of Alphabets

Character	Represe-ntation	Character	Represe-ntation
A	7	N	7851
B	74	O	751
C	78	P	7841
D	785	Q	78451
E	75	R	7451
F	784	S	841
G	7845	T	8451
H	745	U	712
I	84	V	7412
J	845	W	8452
K	71	X	7812
L	741	Y	78512
M	781	Z	7512
:	45	;	41
“	412	!	451
,	4	.	452

Table 2:Mapping of contracted words

Contracted word	Corresponding representation
but	74
can	78
from	748
for	741852
of	74152
the	4182
with	41852
ed	7482
er	74852

In this implementation, 0 is used as delimiter for an alphabet, 3 for word and 6 for the end of sentence. A trie is created for the number sequence from top left to the bottom right with numbers of rows below become the decedents of numbers of rows above. The alphabet of a corresponding number sequence is stored as a leaf. This trie is pre created and used for matching and recognition of Braille alphabets.

Matching of the number sequence is done as the numbers are typed and the corresponding alphabet is displayed when a delimiter is encountered. A voice corresponding to the alphabet is delivered as a feedback to the user. A beep is sound in case of an error. This enables the person to rectify the alphabet immediately.

3.2. Conversion of scanned Braille document

In this method the Braille document is scanned and taken as input, which by a sequence of steps is converted to appropriate text. The scanned document has to be enhanced to identify the dots clearly. The dots are extracted using horizontal and vertical profiling. The Braille cells are

identified and converted to binary sequence. The binary sequence is then mapped to the corresponding alphabets or contracted words. These are stored in a text file and given as input to the voice synthesizer. The basic block diagram is shown in figure 6.

Utmost care is taken to ensure that unwanted noise or redundant information is not introduced at the time of scanning. The scanned image is then converted to gray scale image.

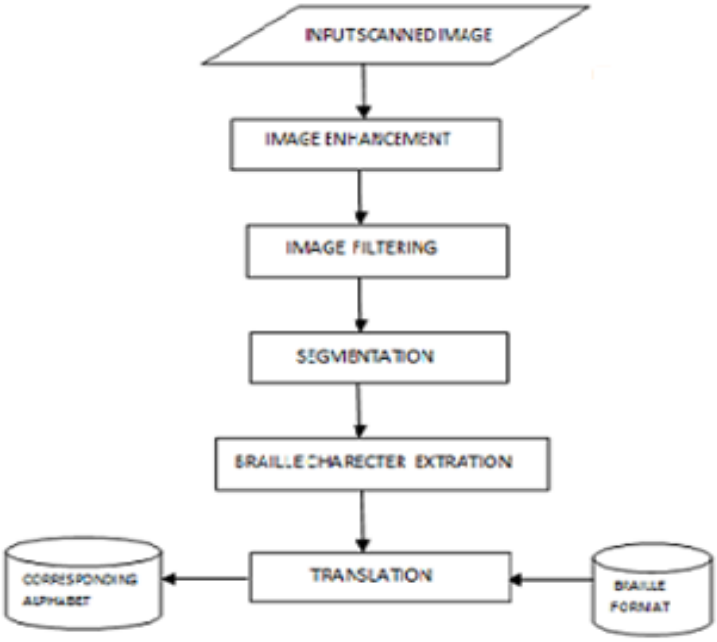


Fig 6.Block Diagram of Proposed Technique

3.2.1. Image enhancement

Due to scanning, the dots in the Braille document cannot be distinguished clearly from the background. Hence various pre processing techniques are applied on the scanned image in order to enhance the dots and to suppress the noise. The dots appear as a darker shade of the back ground color and hence these intensity ranges are identified from the Histogram and enhanced in order to identify the dots. Piece wise enhancement techniques such as contrast stretching, intensity stretching were used for enhancing the dots. These techniques could be represented as S=T(r), where S is the grey level after modification T is the enhancement function used and r is grey level before enhancement.

Contrast stretching is the process that expands the range of the intensity levels in an image as shown in figure 7.This is used to enhance the slightly dark dots from the background. The limits over which image intensity values will be extended are decided from the histogram of the input image.

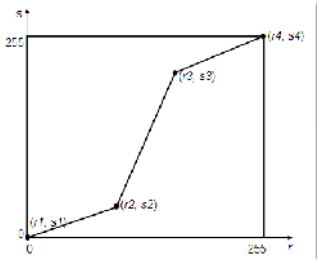


Fig7: Contrast Stretching

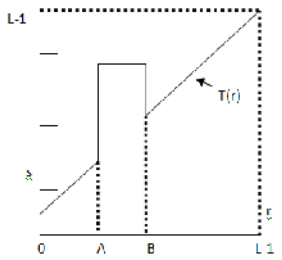


Fig8: Intensity Adjustment

Another level enhancement is done to the dots using Intensity adjustment. This is an image enhancement technique that maps the intensity values of an image to a new range as shown in figure 8. This transformation highlights intensity ranges [A, B] and preserves all other intensity levels.

3.2.2. Image filtering

To remove the unwanted noisy dots present in the scanned documents, the image is smoothened using Gaussian filter and then subjected to morphological opening using a disk shaped structure element B as given in the equation 1

$$A \circ B = (A \ominus B) \oplus B \tag{1}$$

Where $\ominus$ denotes erosion.

An edge detected binary image is obtained using a Prewitt filter. The Prewitt operator uses two 3×3 kernels which are convolved with the image A, to calculate approximations of two derivatives - one for horizontal changes, G_x, and one for vertical, G_y.

$$\mathbf{G_x} = \begin{bmatrix} -1 & 0 & +1 \\ -1 & 0 & +1 \\ -1 & 0 & +1 \end{bmatrix} * \mathbf{A} \quad \text{and} \quad \mathbf{G_y} = \begin{bmatrix} +1 & +1 & +1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} * \mathbf{A} \tag{2}$$

Where * denotes convolution.

The resulting gradient approximations can be combined to give the gradient magnitude, using Eq (3).When magnitude is greater than the threshold T, it is identified as an edge.

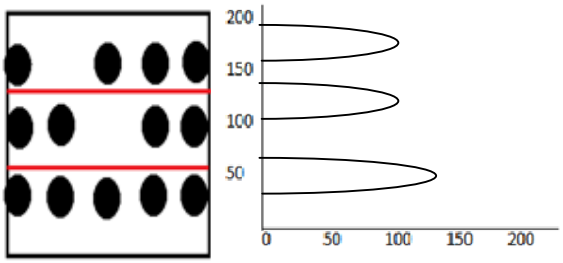
$$G = \left(G_x^2 + G_y^2\right)^{0.5} \tag{3}$$

The edges mostly correspond to the dots of the Braille cells. The border of scanned document and the stapler pin information if any present in the document are removed through image cropping.

3.2.3. Segmenting the Braille Cells

In order to simplify the process of Braille character extraction, the image is first segmented into lines and then into Braille cells. Each cell is further partitioned into binary dot patterns. These are achieved through Projection profiles and standard Braille measurements. Horizontal profiling is performed on edge detected image and zero profile indicates the absence of dots and hence the

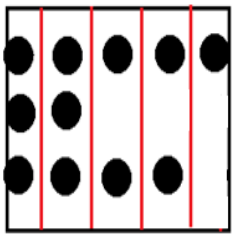
line as shown in Fig 9. Among many such lines, the first line from the top that is closer to the dots is taken as reference. The standard vertical distance between two Braille cells is used to draw the remaining lines where the X projection is zero. This procedure is repeated till the end of the document.



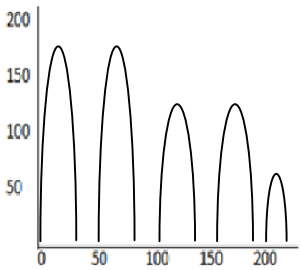
(a) Image (b) Projection profile

Fig9:Horizontal projection profiling

After extracting horizontal lines of the Braille cell, a vertical profiling is performed. Zero profiles indicate the vertical lines as shown in Fig 10. Among many such lines the leftmost one that is closer to the dots is taken as reference. The standard horizontal distance between two Braille cells is used to draw the remaining lines where the Y projection is zero. This procedure is repeated till the end of the document. This segments the edge image in to Braille cells. Each segmented cell is divided into 3 x 2 grids using the standard Braille distance between two dots in a cell.



(a)Image



(b): Projection profile

Fig10: Vertical Projection Profile

3.2.4. Extraction of Text from pattern vector

A Binary pattern vector for each Braille cell is generated. A vector has a length of 6 each correspond to a dot in the Braille cell. The presence of dot is identified after counting the number of white pixels in each grid of a cell and checking whether it satisfies the threshold criterion. '1' indicates that dot is present and '0' indicates that dot is absent in that particular position. This string of bits for the sequence of Braille alphabets is written into a file. A sequence of 6 bits are read from the file and converted to the number sequence and subsequently into the alphabet using the trie structure as discussed in 3.1. If the six bits of the string are 0's, it generates a space. These alphabets are stored in a text file for further processing. Natural Reader [22] is called for reading the converted English text. For Tamil and Hindi, the sequences of 6 bits are taken and the corresponding Unicodes are generated using its pre built mapping table. These Unicodes are stored in the file. The obtained file is converted to the corresponding Tamil and Hindi text. eSpeak[25] is used to read the converted Tamil and Hindi text.

4. EXPERIMENTAL ANALYSIS

The data set includes 20 Braille documents among which 10 are Grade 2 English documents, 5 sheets are Hindi and 5 Tamil sheets. Grade 2 includes contractions which are explained in section 1. Fig11 shows the scanned Braille document. Fig12 shows the edge detected image without applying any enhancement or noise removal techniques. Fig13 shows the edge detected image after applying image enhancement and cropping as explained in section 3.2.

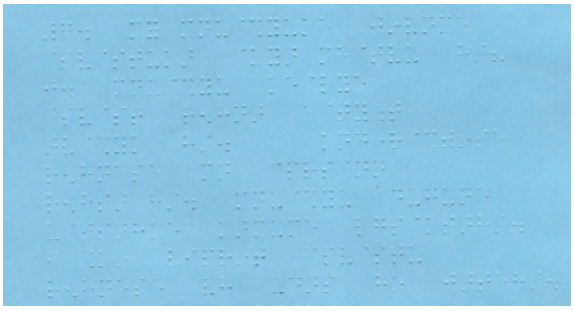


Fig11. Scanned Braille document

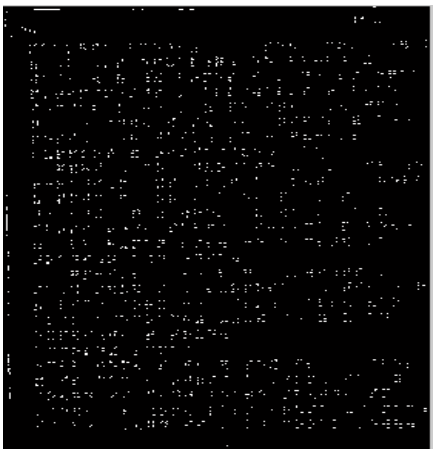


Fig12. Edge detected image with noise

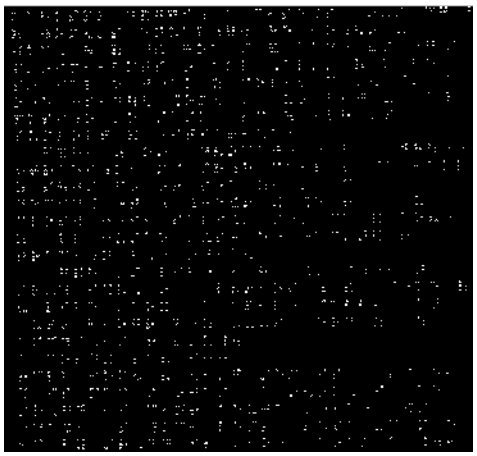


Fig13. Edge detected image without noise

Fig14 shows the image with each line of the document separated by red lines after performing the horizontal profiling. The sample corresponds to Grade2 Braille document. Fig15 shows the image after horizontal and vertical profiling with boxes drawn for each Braille cell. Fig16 focuses on a particular cell after drawing a grid for extracting the dots.

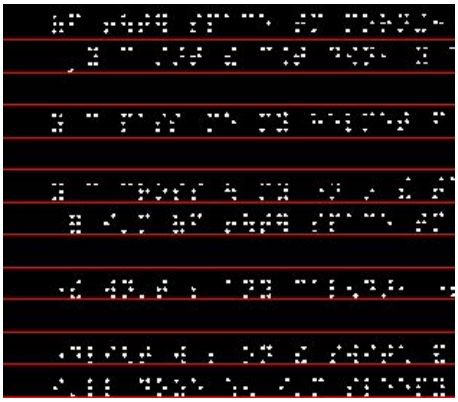


Fig14. Image after horizontal profiling

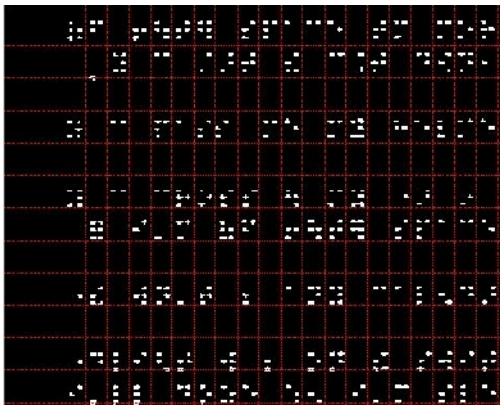


Fig15.Image after horizontal and vertical profiling

Fig16. Extraction of each dot in a cell

After the extraction of the dots, a binary sequence is generated as shown in Fig17 where value '1' shows dot is present and '0' shows dot is absent in a cell. The English alphabet sequence mapped from the binary sequence is shown in Fig18. The mapped alphabets for Tamil and Hindi for two other documents are shown in Fig19, Fig 20 respectively. The errors occurring due to wrong punching of Braille characters are indicated in blue colour box. The errors occurring while combining the Unicodes are indicated in red colour box. When a short notation of a word is used, those are not expanded by the system. For example 'tm' is a short notation used for 'tomorrow'. These are indicated in green box. In Hindi document, appropriate spacing was not punched in the input Braille document and hence the words appear consecutively.

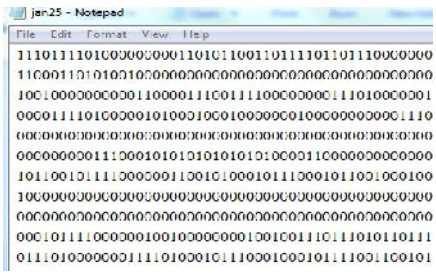


Fig17. Binary sequence for English

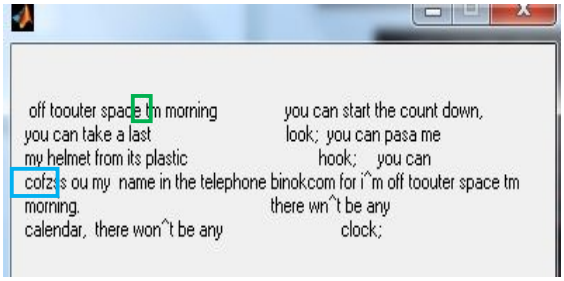


Fig18. Final mapped text for English

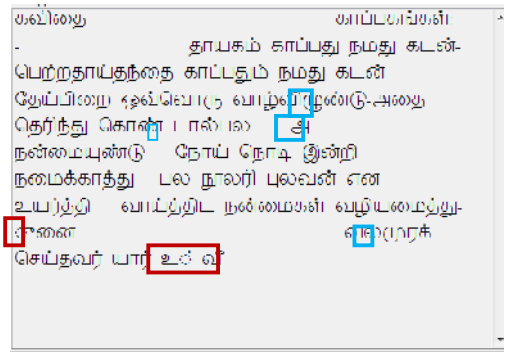


Fig19. Final mapped text for Tamil

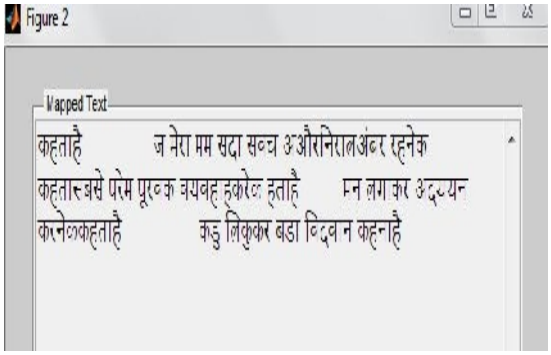


Fig20. Final mapped text for Hindi

To evaluate the performance of the system, each Braille document is decoded manually and compared with the system results. The effect of various enhancement techniques on the system performance is tabulated for English document in Table3. The table shows the percentage of accuracy of the words correctly identified after mapping. In the table CS represents contrast stretching, IS represents intensity stretching and MO represents morphological operations. The sequence of image enhancement techniques when applied on the image also influences the accuracy of the words mapped. The accuracy for few sequences is shown in column 2 through 4. The accuracy is found to be high for the sequence CS, IS, and MO. The accuracy drops drastically when the order of enhancement techniques is changed. These are illustrated in the table. The number of words correctly mapped and the percentage of accuracy for the English(E#), Tamil(T#) and Hindi(H#) documents are shown in Table 4. In this table, TW represents the total number of words present in the Braille document and C represents the number of words correctly mapped by the system, final column specifies the percentage of accuracy in mapping. Confusion occurs when two different symbols have same Braille representation. For example in English, ‘,’ and ‘ea’ ; ’be’ and ’;’ in Tamil letter ஞ and symbol ‘:’ has same braille representation. Accuracy drops when such confusion occurs. In Hindi the percentage of accuracy is dropped because of the manual mistakes done while punching the Braille document.

Table 3: Analysis of enhancement Techniques.

	CS	CS,IS	CS,IS,MO	MO,CS,IS
Doc 1	75%	86%	97%	32%
Doc 2	72%	88%	97%	34%
Doc 3	80%	85%	98%	36%
Doc 4	71%	80%	97%	34%
Doc 5	75%	86%	93%	24%
Doc 6	73%	83%	97%	39%
Doc 7	77%	82%	96%	26%
Doc 8	70%	89%	95%	33%
Doc 9	80%	89%	96%	32%
Doc 10	78%	88%	91%	29%

Table 4: Accuracy of text conversion.

	TW	C	Accuracy(%)
E 1	343	341	99.4
E 2	350	346	98.2
E 3	425	425	100
E4	327	320	97.8
E 5	402	398	99.0
E 6	250	247	98.8
E 7	361	354	98.0
E 8	393	392	99.7
E 9	285	283	99.2
E 10	324	324	100
T 1	378	375	99.2
T 2	405	399	98.5
T 3	290	287	98.9
T 4	318	318	100
T 5	328	326	99.3
H 1	345	341	98.8
H 2	323	320	99.0
H 3	298	297	99.6
H 4	276	272	98.5
H 5	354	351	99.1

5. LIMITATIONS and ADVANTAGES

Since the standard Braille dimensions are used for the segmentation of the Braille cells, the document has to be free from tilt and has to be aligned with the edge of the scanner. This poses a major limitation to the system. The presence of the unnecessary dots or noises whose size is comparable to that of the Braille dots during scanning is difficult to remove during pre processing and hence affects the accuracy of the converted text.

It involves very less intervention of the user and helps to serve the need of large number of people using a single document. It helps resource teachers in Inclusive Education, who do not know Braille. Simplifies making of copies of old Braille books for which only one copy is available as it saves the labour of preparing the same again. Since the availability of Braille document is also limited, scanning the document also help in preserving the existing documents.

6. CONCLUSION

This paper focuses on the conversion of scanned Braille documents to corresponding text in English language and Indian languages namely Hindi and Tamil. After identifying the start of the Braille text, the lines and subsequently the Braille cell are segmented. Grids are drawn based on the standard measurement of the Braille cells and the dots are extracted. Braille has a standard pattern of alphabets and only the mapping differs from language to language. Using appropriate mapping for each language the alphabets are identified and stored as text. These texts are read out by voice synthesizer. The extraction of the dots was affected when they were not confined to the standard measurement and due to the presence of noise during scanning. Mapping errors occurred when the Braille has similar representation for the alphabet and the punctuation. These are

eliminated to some extent using simple rules governing the language. The mapping errors are predominant for Grade 2 English documents. The voice synthesizer used for speaking the Native languages had a poor pronunciation. The paper could be extended for Grade 3 English documents and the voice synthesizer for Hindi and Tamil could be customized.

REFERENCES

[1] Saad D. Al-Shamma and Sami Fathi, "Arabic Braille Recognition and Transcription into Text and Voice", 2010 5th Cairo International Biomedical Engineering Conference Cairo, Egypt, December 16-18, 2010, Pages 227-231

[2] AbdulMalik S. Al-Salman, "A Bi-directional Bi-Lingual Translation Braille-Text System", J. King Saud University, Vol. 20, Comp. & Info. Sci., pp. 13-29,Riyadh(1428H./2008).

[3] Charanya C, Kalpana S and Nithya R, "Real time Braille recognition with sonic feedback", Intel India Research Challenge 2007

[4] Er.Sheilly Padda, Er. Nidhi, Ms. Rupinderdeep Kaur,"A Step towards Making an Effective Text to speech Conversion System",International. Vol. 2, Issue 2,Mar-Apr 2012, pp.1242-1244

[5] http://www.braillemaster.com

[6] Manzeet Singh ,Parteek Bhatia,"Automated conversion of English and Hindi text to Braille representation",International Journal of Computer Applications, vol. 4, issue 6, pp. 25-29, year 2010

[7] Xuan Zhang, Cesar Ortega-Sanchez and Iain Murray, "A System for Fast Text-to-Braille Translation Based on FPGAs",SPL2007 - III Southern Conference on Programmable Logic, Mar del Plata, Argentina, February 26-28, 2007

[8] Minghu Jiang etal, "Braille to print translations of Chinese",Information and Software Technology 44 (2002) 91-100

[9] Trends And Technologies In Optical Braille Recognition by AbdulMalik S. Al-Salman, Yosef AlOhali, and Layla O. Al-Abdulkarim, 3'rd Int. Conf. on Information Technology,May 2007,Jordan.

[10] AbdulMalik,S. Al-Salman,"A Bi-directional Bi-Lingual Translation Braille-Text System", Journal of King Saud University - Computer and Information Sciences Volume 20, 2008, Pages 13–29

[11] AbdulMalik Al-Salman, Yosef AlOhali, Mohammed AlKanhal, and Abdullah AlRajih,"An Arabic Optical Braille Recognition System",ICTA'07, April 12-14, Hammamet, Tunisia

[12] Roman Graf, Reinhold Huber-Mörk, "A Braille Conversion Service Using GPU and Human Interaction by Computer Vision", Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES 2011), 2011, 190-193.

[13] Lisa Wong,Waleed Abdulla,Stephan Hussmann,"A Software Algorithm Prototype for Optical Recognition of Embossed Braille", Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on 23-26 Aug. 2004, 586- 589 Vol.2

[14] R.T. Ritchings, A. Antonacopoulos and D. Drakopoulos,"ANALYSIS OF SCANNED BRAILLE DOCUMENTS",In the book: Document Analysis Systems, A. Dengel and A.L. Spitz (eds.), World Scientific Publishing Co,1995, pp. 413-421

[15] Rawan Ismail Zaghloul,Tomader Jameel Bani-Ata,"Braille Recognition System – With a Case Study Arabic Braille Documents",European Journal of Scientific Research ISSN 1450-216X Vol.62 No.1 (2011), pp. 116-122

[16] C M Ng, Vincent Ng, Y Lau,"Regular Feature Extraction for Recognition of Braille", Computational Intelligence and Multimedia Applications, 1999. ICCIMA '99. Proceedings. Third International Conference, pages 302-306

[17] Zainab I. Authman, Zamen F.Jebr, "Arabic Braille scripts recognition and translation using image processing techniques", Journal: Journal of College of Education, Year: 2012 Volume: 2 Issue: 3 Pages: 18-26, Publisher: Thi-Qar University

[18] C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran,"A Bilingual OCR for Hindi-Telugu Documents and its Applications", Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on 3-6 Aug. 2003, Page(s): 408 - 412 vol.1

[19] Pradeep Manohar and Aparajit Parthasarathy,"An Innovative Braille System Keyboard for the Visually Impaired", In Proceedings of UKSIM. 2009, 559-562.

[20] Jan mennues etal.,"Optical Recognition of Braille writing Using Standard Equipment", IEEE TRANSACTIONS ON REHABILITATION ENGINEERING, VOL. 2, NO. 4, DECEMBER 1994

[21]Sparsha: A Comprehensive Indian Language Toolset for the Blind by Anirban Lahiri, Satya Jyoti Chattopadhyay, Anupam Basu, ASSETS 2005 - The Seventh International ACM SIGACCESS Conference on Computers and Accessibility, October 9-12, 2005, Pages 114-120

[22]http://www.softpedia.com/get/Multimedia/Audio/Other-AUDIO-Tools/Free-NaturalReader.shtml

[23]Paul Blenkhorn,"System For Converting Braille Into Print",IEEE TRANSACTIONS ON REHABILITATION ENGINEERING, VOL. 3, NO. 2, JUNE 1995

[24]Li Nian-feng,Wang Li-rong,"A kind of Braille paper automatic marking system", Mechatronic Science, Electric Engineering and Computer (MEC), 2011 International Conference on 19-22 Aug. 2011 Page(s): 664- 667

[25]http://www.softpedia.com/progDownload/eSpeak-Download-75752.html

electronics

MDPI

Article

Smart Glass System Using Deep Learning for the Blind and Visually Impaired

Mukhriddin Mukhiddinov and Jinsoo Cho *

Department of Computer Engineering, Gachon University, Sujeong-gu, Seongnam-si 13120, Korea; mukhiddinov18@gachon.ac.kr
* Correspondence: jscho@gachon.ac.kr

Abstract: Individuals suffering from visual impairments and blindness encounter difficulties in moving independently and overcoming various problems in their routine lives. As a solution, artificial intelligence and computer vision approaches facilitate blind and visually impaired (BVI) people in fulfilling their primary activities without much dependency on other people. Smart glasses are a potential assistive technology for BVI people to aid in individual travel and provide social comfort and safety. However, practically, the BVI are unable move alone, particularly in dark scenes and at night. In this study we propose a smart glass system for BVI people, employing computer vision techniques and deep learning models, audio feedback, and tactile graphics to facilitate independent movement in a night-time environment. The system is divided into four models: a low-light image enhancement model, an object recognition and audio feedback model, a salient object detection model, and a text-to-speech and tactile graphics generation model. Thus, this system was developed to assist in the following manner: (1) enhancing the contrast of images under low-light conditions employing a two-branch exposure-fusion network; (2) guiding users with audio feedback using a transformer encoder–decoder object detection model that can recognize 133 categories of sound, such as people, animals, cars, etc., and (3) accessing visual information using salient object extraction, text recognition, and refreshable tactile display. We evaluated the performance of the system and achieved competitive performance on the challenging Low-Light and ExDark datasets.

Keywords: smart glasses; artificial intelligence; blind and visually impaired; deep learning; low-light images; assistive technologies; object detection; refreshable tactile display

Citation: Mukhiddinov, M.; Cho, J. Smart Glass System Using Deep Learning for the Blind and Visually Impaired. *Electronics* **2021**, *10*, 2756. https://doi.org/10.3390/electronics10222756

Academic Editor: Amir Mosavi

Received: 27 September 2021
Accepted: 8 November 2021
Published: 11 November 2021

1. Introduction

In the modern era of information and communication technology, the lifestyle and independent movement of blind and visually impaired people is among the most significant issues in society that need to be addressed. Governments and various specialized organizations have enacted many laws and standards to support people with visual disabilities and have organized essential infrastructure for them. According to the World Health Organization, at least 2.2 billion people worldwide suffer from vision impairment or blindness, of whom at least 1 billion have a vision impairment that could have been prevented or is yet to be addressed in 2020 [1]. Vision impairment or blindness may be caused by several reasons, such as, cataract (94 million), unaddressed refractive error (88.4 million), glaucoma (7.7 million), corneal opacities (4.2 million), diabetic retinopathy (3.9 million), trachoma (2 million), and others [1]. The primary problems that blind and visually impaired (BVI) people encounter in their routine lives involve action and environmental awareness. Several solutions exist to such problems, employing navigation and object recognition methods. However, the most effective navigation methods, such as a cane, trained guide dogs, and smartphone applications suffer from certain drawbacks; for example, a cane is ineffectual over long distances, crowded places, and cannot provide

information regarding dangerous objects or car traffic when crossing the street, whereas training of guide dogs is cumbersome and expensive, and dogs require special attention when caring for them. Further, although smartphone applications such as voice assistance and navigation maps for BVI people are evolving rapidly, proper and complete use is still low.

Recent advancements in embedded systems and artificial intelligence have had a significant impact on the field of wearable assistive technologies for the visually impaired, and consequently several devices have been developed and placed on the market. Assistive systems exist to aid BVI people with navigation and daily activities, such as distinguishing banknotes [2,3], crossing a road [4–6], video media accessibility [7,8], image sonification for navigation [9,10], recognizing people [11–13], recognizing private visual information [14], selecting clothing [15], and navigating both outdoors and indoors [16–18]. Daescu et al. [13] proposed a face recognition system with smart glasses and used a server-based deep learning face recognition model. In this system, they used client–server architecture to reduce power consumption and computational time. Joshi et al. [17] introduced an assistive device to recognize different objects based on a deep learning model, and a distance-measuring sensor was combined to make the device more complete by identifying barriers while traveling from one place to another.

However, owing to the low-light environment and the lack of light at night, problems such as recognizing people and objects as well as providing accurate information to the BVI people are prevalent. Moreover, the visually impaired face serious challenges, particularly when walking in public places, where simple actions such as avoiding obstacles, crossing the street, and using public transportation can pose significant risks and challenges. Such challenges endanger the safe and confident independent action of BVIs and limit their ability to adapt and experience liberty in public life.

Among the wearable assistive devices that are considered the most comfortable and useful for BVI are smart glasses, which can achieve the original goal of providing clearer vision while operating similarly to a computer. In the nearly nine years since Google announced its assistive device called "Google Glass" for BVI in 2013, many companies such as Epson, Sony, Microsoft, Envision, eSight, NuEyes, Oxsight, and OrCam have started producing smart glasses with various degrees of usability. Most such glasses have an embedded operating system and support Wi-Fi or Bluetooth for wireless communication with external devices to aid in the exploration and receiving of information in real time through the Internet, in addition to the built-in camera. Further, a touch sensor or voice recognition method may be employed as an interface to facilitate interaction between users and the smart glasses they wear. Moreover, images or video data of the surrounding environment can be obtained in real time by mounting a camera in front of the device and using computer vision methods. In the following Table 1, we compared the performance and parameters of the proposed system and other commercially available smart glass solutions for BVIs.

Table 1. The performance comparison of the proposed system and commercially available smart glasses.

Smart Glasses	Target Users	Object Recognition	Text Recognition	Independent	Tactile Graphics	Walking Night-Time	Battery Capacity
eSight 4 [19]	Low vision	No	Yes	Yes	No	No	2 h
NuEyes Pro [20]	Low vision	No	Yes	Yes	No	No	3.5 h
OrCam My Eye [21]	BVI	Yes	Yes	Yes	No	No	NA
Oxsight [22]	VI	Yes	Yes	Yes	No	No	2 h
Oton glass [23]	Low vision	No	Yes	Yes	No	No	NA
AngleEye [24]	Low vision	Yes	Yes	Yes	No	No	2 h
EyeSynth [25]	BVI	No	No	Yes	No	No	8 h
Envision [26]	BVI	Yes	Yes	Yes	No	No	5.5 h
Our System	BVI	Yes	Yes	Yes	Yes	Yes	6 h

Recently, researchers published review papers by analyzing the features of wearable assistive technologies for the BVI. Hu et al. [27] reviewed various assistive devices such as glasses, canes, gloves, and hats by analyzing behavior, structure, function, principle, context, and state of the wearable assistance system. Their analysis includes various assistive devices and 14 assistive glass research works, and 6 assistive glasses available in the market. Based on the analysis, various assistive devices can only perform on a restricted spatial scale due to their insufficient sensors and feedback methods [27]. In 2020, another survey paper on assistive technologies for BVI was published by Manjari et al. [28]. Assistive devices which were developed until 2019 were gathered and the advantages and limitations of those devices were discussed. In this year, Gupta et al. [29] studied and explored the existing assistive devices which assisted in day-to-day tasks with a simple and wearable design to give a sufficient user experience for BVI. Their findings were as follows: many assistive devices are focused on only one aspect of the problem, making it challenging for users to have a better experience; assistive devices are very heavy on a user's pocket in comparison with the features they provide, making them quite difficult to afford [29]. El-Taher et al. [30] presented a broad analysis of research relevant to assistive outdoor navigation along with commercial and non-commercial navigation applications for BVI from 2015 until 2020. One of their findings related to our smart glass system is that camera-based systems are affected by illumination and weather conditions; however, they provide more features around barriers such as shape and color.

The use of machine learning and object detection and recognition methods based on deep learning models ensures that the results obtained based on the received data from the users are reliable. In addition, before applying object recognition, the use of preprocessing methods such as contrast enhancement and noise removal is crucial, particularly in the case of low-light and dark images. Researchers have designed several approaches for the development of smart glass systems [31–34] to aid BVI people with navigation. However, these systems have major limitations. First, most of them do not employ recently developed computer vision methods, such as deep learning, and thus, the results are not efficient and less reliable. Second, they were developed to assist BVI people in crossing a road using pedestrian signals and zebra-crossing or bollard detection and recognition methods. Third, they were created considering daytime conditions and a good environment, while night-time and low-light environments were ignored. However, the effective use of cutting-edge deep learning in low-light image enhancement and object detection and recognition methods can improve awareness of surroundings and assist the confident travel of BVI people during night-time.

In this study, we proposed a smart glass system that employs computer vision techniques and deep learning models for BVI people. It was designed based on the fact that BVI people have a desire to travel at any time of the day. The proposed system can navigate BVI people even in night-time environments and comprises three parts: (1) low-light image enhancement using a two-branch exposure-fusion structure based on a convolutional neural network (CNN) to overcome noise and image enhancement limitations [35]; (2) object recognition based on a deep encoder–decoder structure using transformers [36] to help the visually impaired navigate with audio guidance; (3) salient object detection and tactile graphics generation using a two-level nested U-structure network [37] and results of our previous work [38]. Figure 1 shows the structure of a wearable navigation system with a smart glass and a refreshable tactile display. As presented in Figure 1, the main focus of this paper is marked with a red outline, and the methods in the computer vision module are explained in detail.

The primary contribution of the proposed work is as follows:

- A fully automated smart glass system was developed for BVI people to assist in the cognition process of surrounding objects during night-time. To the best of our knowledge, existing smart glass systems do not support walking in the night-time and a low-light noise environment and cannot handle night-time problems (Table 1).

- It provides users with information regarding surrounding objects through real-time audio output. In addition, it provides additional features for users to perceive salient objects using their sense of touch through a refreshable tactile display.
- The proposed system has several advantages compared to the previously developed systems; that is, the use of deep learning models in computer vision methods, and not being limited to only object detection, and global positioning system (GPS) tracking methods using basic sensors. It has four main deep learning models: low-light image enhancement, object detection, salient object extraction, and text recognition.

The remainder of this paper is organized as follows. In Section 2, we review the literature on smart glass systems and object detection and recognition. Section 3 explores the proposed system. Sections 4 and 5 discus the experimental results and highlights certain limitations of the proposed system respectively. Finally, the conclusions are presented in Section 6 including a summary of our findings and the scope for future work.

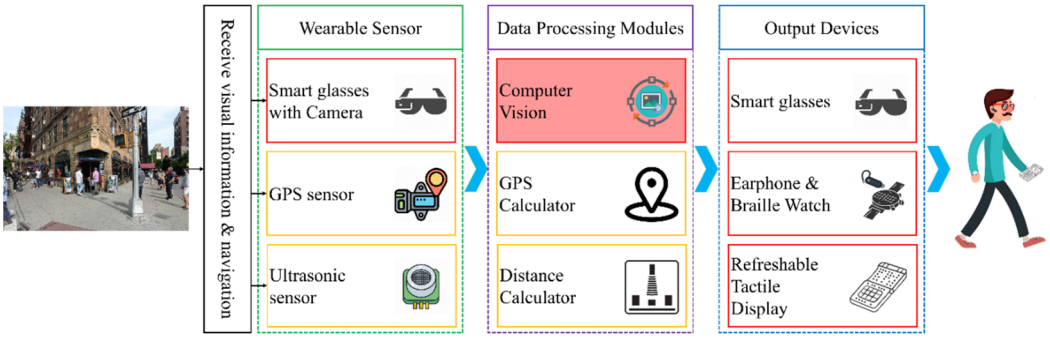


Figure 1. The structure of a wearable navigation system with smart glass and refreshable tactile display.

2. Related Works

In this section, we review studies conducted in the field of smart glass systems and object recognition. Wearable assistance systems have been developed as one of the most convenient and efficient solutions for BVI people to facilitate independent movement and performance of daily personal tasks. Smart glass systems have been employed in many fields such as health care, assisting people with visual disabilities, computer science, social science, education, service, industry, agriculture, and sports. In this literature review, we highlight the beneficial aspects of BVI people.

2.1. Smart Glass System for BVI People

One of the most important and significant tasks for BVI people is to recognize the face and identity information of relatives and friends. Daescu et al. [13] created a face recognition system that can receive facial images captured via the camera of smart glass based on commands from the user, process the result on the server, and thereafter return the result via audio. The system is designed as a client–server architecture, with a pair of cellphones, smart glasses, and a back-end server employed to implement face recognition using deep CNN models such as FaceNet and Inception-ResNet. However, this face recognition system needs to retrain to recognize new faces that are not available on the server, thereby requiring increased time to function. Mandal et al. [39] focused on the ability of recognition of faces under various lighting conditions and face poses and developed a wearable face recognition system based on Google Glasses and subclass discriminant analysis to achieve within-subclass discriminant analysis. However, this system suffers from a familiar problem; that is, although it correctly recognized the faces of 88 subjects, the model had to be retrained for new faces that were not in the initial dataset.

Further, the high price of existing commercial assistive technologies induces immense financial stress to most BVI people in developing countries and even developed countries. To solve this problem, Chen et al. [40] introduced a smart wearable system that performs

object recognition from input video frames. Their system is also built on client–server architecture, and the main image processing processes are performed on the server side, while the client side only captures images and feeds the results back to the users. As a result, the processor of the system need not employ high-priced tools, significantly reducing the cost. They used Raspberry Pi, a micro camera, and an infrared and ultrasonic sensor as the local unit, connected to the Baidu cloud server via Wi-Fi or 4G network. Furthermore, the image processing algorithm operating on the cloud server guaranteed speed and accuracy, which coupled with capturing points of interest mechanism reduced the power consumption. Ugulino and Fuks [41] described cocreation workshops and wearables prototyped by groups of BVI users, designers, mobility instructors, and computer engineering students. The group merges verbalized warnings with audio feedback and haptics to assist BVI people in recognizing landmarks. The recognition of landmarks is a necessary experience that is challenging for spatial representation and cognitive mapping. Kumar et al. [42] proposed a smart glass system to recognize objects and obstacles. It was designed with Raspberry Pi, ultrasonic sensors, mini camera, earphones, buzzer, power source, and controlled via a button to acquire photos of the surroundings concerning the user position. The primary purpose of the system was to recognize the surrounding objects using Tensorflow models and consequently alert the blind regarding collisions with obstacles via audio using ultrasonic sensors.

Traveling in large open areas and reaching the desired point poses various problems for the visually impaired because there are no tactile pavers and braille guides at such places. Consequently, Fiannaca et al. [43] proposed a navigation aid that assists BVI users using Google Glass to travel in large open areas. Their system provides secure navigation toward salient landmarks such as doors, stairs, hallway intersections, floor transitions, and water coolers by providing audio feedback to guide the BVI user towards landmarks. However, experimental results indicated that blind people typically hold the cane in their right hand to aid in navigation, which causes problems in commanding the touchpad of the smart glass using the right hand. The touchpad should be on the left side to provide a more efficient interaction with smart glass while using a cane and smart glass in parallel.

Another interesting research approach is to solve the eye contact problem of blind people in a community to facilitate conversations via eye contact with their sighted friends or partners. This problem causes feelings of social isolation and low confidence in conversations. A social glass system and tactile wrist band were implemented by Qiu et al. [44]. These two assistive devices are worn by BVI people and they assisted them in establishing eye contact and tactile feedback when eye contact was observed between blind and sighted people. Lee et al. [45] presented a concept solution to assist visually impaired people in acquiring visual information regarding pedestrians in their environment. A client and server were included in the concept solution. A server component analyzed the visual data and recognized a pedestrian based on photographs captured by the client. Face recognition, gender, age, distance calculations, and head pose are among the features available on the server. The client acquired photos and provided audio feedback to users using text-to-speech (TTS).

Furthermore, using only ultrasonic sensors in smart glass systems has also received much attention from researchers [46–48]. Hiroto and Katsumi [46] introduced a walking support system that has a glass-type wearable assistive device with an ultrasonic obstacle sensor and a pair of bone conduction earphones. Adegoke et al. [47] proposed a wearable eyeglass with an ultrasonic sensor to assist BVI people in safe navigation while avoiding objects that may be encountered, fixed, or movable, hence eliminating any potential accidents. Their system detects objects at a distance of 3–5 m, and the controller quickly alerts the user through voice feedback. However, no camera is installed to analyze the surroundings of the BVI people.

To solve the above-mentioned limitations and problems, the proposed system applied four deep learning models: low-light image enhancement, object detection, salient object extraction, and text recognition, and used the client–server architecture. The main advantages

of the proposed system over other existing systems is supporting tactile graphics generation and walking in night-time environment. Note that other existing works [13,40,45] also used a client–server architecture and increased smart glass's battery life and decreased data processing time.

2.2. Object Detection and Recognition Models

In recent years, artificial intelligence and deep learning approaches are rapidly entering all areas, including autonomous vehicle systems [49,50], robotics, space exploration, medicine, pet and animal monitoring systems [51], and areas that start with the word smart, such as smart city, smart home, smart agriculture, etc. Computer vision and artificial intelligence methods play a key role in the development of smart glass systems. It is not possible to build a smart glass system without computer vision methods such as object detection and recognition methods because the input data is an image or a video. Object detection and recognition has garnered the attention of researchers, and numerous new approaches are being developed every year. To reduce the review areas, we analyzed lightweight object detection and recognition models designed for embedded systems.

In 2016, Iandola et al. [52] designed three primary mechanisms to squeeze CNN networks and named SqueezeNet: (1) 3×3 filters were replaced with 1×1 filters; (2) the number of input channels was reduced to 3×3 filters, and (3) the network was downsampled late. These three approaches reduced the number of parameters in a CNN while maximizing the accuracy of the limited parameter sources. Further, the fire module was utilized in SqueezeNet's network architecture, which contained squeeze convolution and expansion layers. The former consists of only 1×1 convolutional filters and is fed into an expanded layer that comprises a mix of 1×1 and 3×3 convolutional filters. The output of the expanded layer is concatenated in the channel dimension such that one layer contains 1×1 convolution filters and 3×3 convolution filters. The model size achieved $50\times$ reduction compared to AlexNet and a size less than 0.5 MB was possible using the deep compression technology. Chollet [53] improved InceptionV3 by replacing a convolution with a depth-wise separable convolution and introduced the Xception model. This depth-wise separable convolution approach has been extensively applied in many other popular models such as MobileNet [54,55], ShuffleNet [56,57], and other network architectures. However, the implementation of depth-wise separable convolution is not sufficiently efficient for deep CNNs.

Mobile deep learning is rapidly expanding. The Tiny-YOLO net for iOS, introduced by Apte et al. [58] in 2017, was developed for mobile devices and tested with a metal GPU for real-time applications with an accuracy approximately similar to the original YOLO. In the same year, Howard et al. [54] built a lightweight deep neural network named MobileNet using depth-wise separable convolution architecture for mobile and embedded systems. This model has inspired researchers and has been used in various applications. In 2018, the MobileNet-SSD network [59], derived from VGG-SSD, was proposed to improve the accuracy of small objects in real-time speed. Further, Wong et al. [60] developed a compact single-shot detection deep CNN based on the remarkable performance of the fire microarchitecture presented in SqueezeNet [52] and the macro architecture introduced in SSD. A tiny SSD is created for real-time embedded systems by reducing the model size and consists of a fire subnetwork stack and optimized SSD-based convolutional feature layers. With the increasing capabilities of processors for mobile and embedded devices, numerous effective mobile deep CNNs for object detection and recognition have been introduced in recent years, such as ShuffleNet [56,57], PeleeNet [61], and EfficientDet [62].

3. The Proposed Smart Glass System

Our goal is to create convenience and opportunities for BVI people to facilitate independent travel during both day and night-time. To achieve this goal, wearable smart glass and a multifunctional system that can capture images through a mini camera and return object recognition results with voice feedback to users are the most effective approaches.

It is also conceivable to perceive visual information by touching the contours of detected salient objects according to the needs of blind people via a refreshable tactile display. The system is required to use deep CNNs to detect objects with high accuracy, and a powerful processor to perform the processes sufficiently fast in real time. Therefore, we introduced client–server architecture that consists of smart glass and a smartphone/tactile pad [63] as a local, and an artificial intelligence server to perform image processing tasks. Hereinafter, for simplicity in the text, a smartphone is written instead of a smartphone/tactile pad. The overall design of the proposed system is illustrated in Figure 2. The local part comprises smart glass and a smartphone and transfers data via a Bluetooth connection. Meanwhile, the artificial intelligence server receives the images from the local, processes them, and returns the result in audio format. Note that, smart glass hardware has a built-in speaker for direct output and earphone port for audio connection to convey returned audio results from smartphone to users.

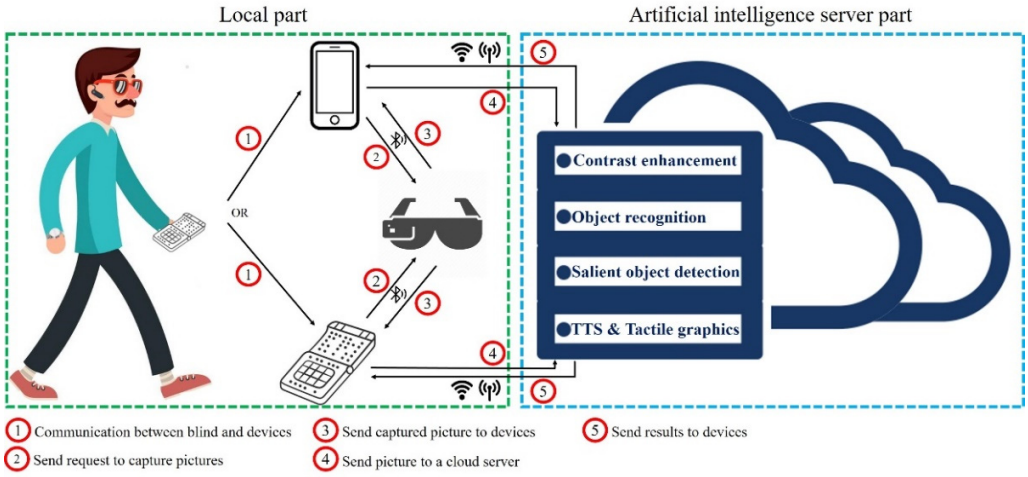


Figure 2. The overall design of the proposed system.

The working of the local part is as follows: first, the user makes a Bluetooth connection between a smart glass and a smartphone. Following this, the user can send a request to the smart glass to capture images, and the smartphone receives the images. In this scenario, the power consumption of smart glasses can be reduced, which is much more efficient than continuous video scanning. Thereafter, the results from the artificial intelligence server are delivered in voice feedback via earphones or speaker or smartphone. Further, tactile pad users can touch and sense the contours of the salient objects. Although lightweight deep CNN models have been introduced recently, we performed object detection and recognition tasks on an artificial intelligence server because the capabilities of the GPUs within wearable assistive devices and smartphones are limited compared to an artificial intelligence server. In addition, this increases the battery life of smart glasses and smartphones because they are used only for capturing images.

The artificial intelligence server part includes four main models: (1) a low-light image enhancement model, (2) an object detection and recognition model, (3) a salient object detection model, and (4) a TTS and tactile graphics generation model. Further, the artificial intelligence server part functions under two modes depending on sunrise and sunset times: daytime and night-time. In the daytime mode, the low-light image enhancement model does not function. The working of the nighttime mode is as follows (Figure 3): first, the system runs a low-light image enhancement model to increase the dark image quality and remove noise after receiving an image from a smartphone. Following the improvement in the image quality, object detection, salient object extraction, and text recognition models are applied to recognize objects, and text-to-speech is conducted. Subsequently, the audio results are returned as an artificial intelligence server response to the request made by the local. If the image is received from the tactile pad with a special title, the salient object

detection model is also performed, and the tactile graphics are also sent with the audio results as a response.

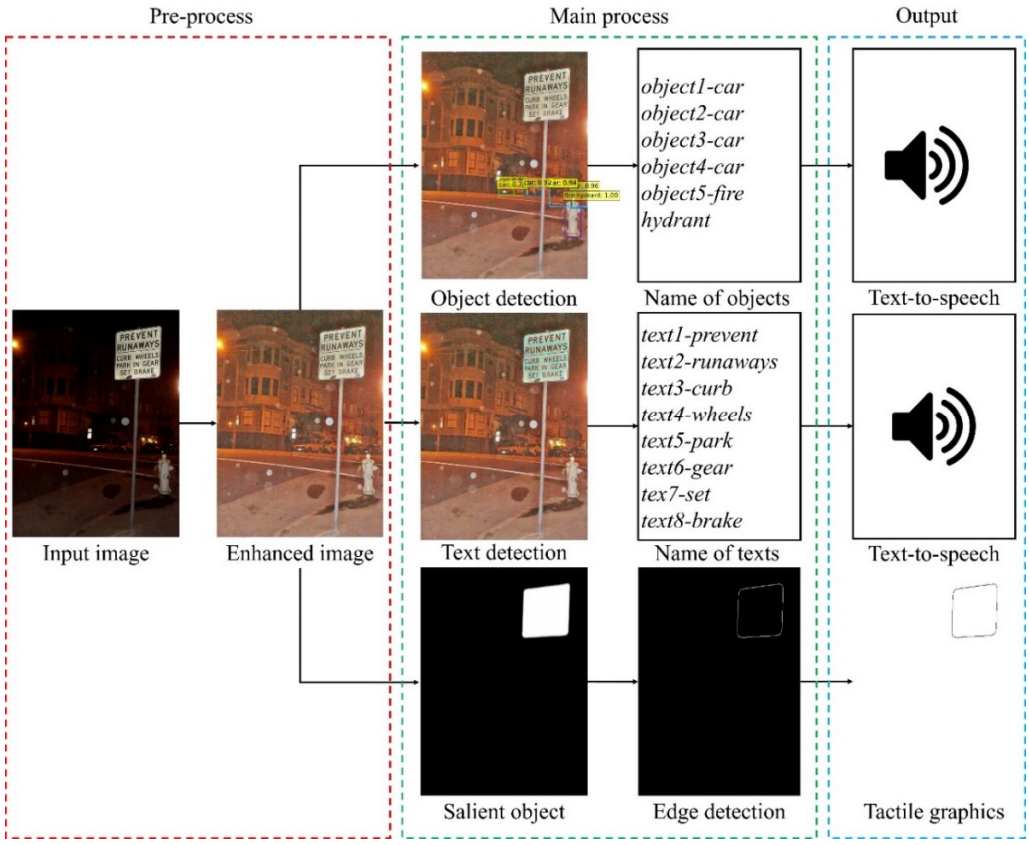


Figure 3. The main steps of the proposed system.

3.1. Low-Light Image Enhancement Model

Low-light images typically have very dark zones, blurred features, and unexpected noise, particularly when compared with well-illuminated images. This can appear when the scene is nearly dark, such as under limited luminance and night-time, or when the cameras are not set correctly. Consequently, such images show low quality owing to unsatisfactory processing of information when creating high-level applications such as object detection, recognition, and tracking owing to poor quality. Thus, this area of research is among the most valuable in computer vision, and has attracted the attention of many researchers because it is of high importance in both low-level and high-level applications such as self-driving, night vision, assistive technologies, and visual surveillance.

The use of a low-light image enhancement model for the BVI to move independently and comfortably in the dark would be an appropriate and effective solution. A low-light image enhancement model based on deep learning has recently achieved high accuracy while removing various noises. Therefore, we used a two-branch exposure-fusion network based on a CNN [35] to realize a low-light image enhancement model. A two-branch exposure-fusion network consists of two stages, wherein a two-branch illumination enhancement framework is applied in the initial step of the low-light improvement procedure, where two different enhancing approaches are employed independently to enhance the potential. A data-driven preprocessing module was presented to relieve the degradation under considerably dark conditions. Subsequently, these two enhancing modules were fed into the fusion module in the second step, which was trained to combine them with a fundamental but effective attention strategy and refining procedure. In Figure 4, we present the overall architecture of a two-branch exposure-fusion network [35]. Lu and Zhang referred to the two branches as -1E and -2E because the upper branch provides

greater support for images in the evaluation set with an exposure level of -1E, while the other branch provides greater support for images with an exposure level of -2E.

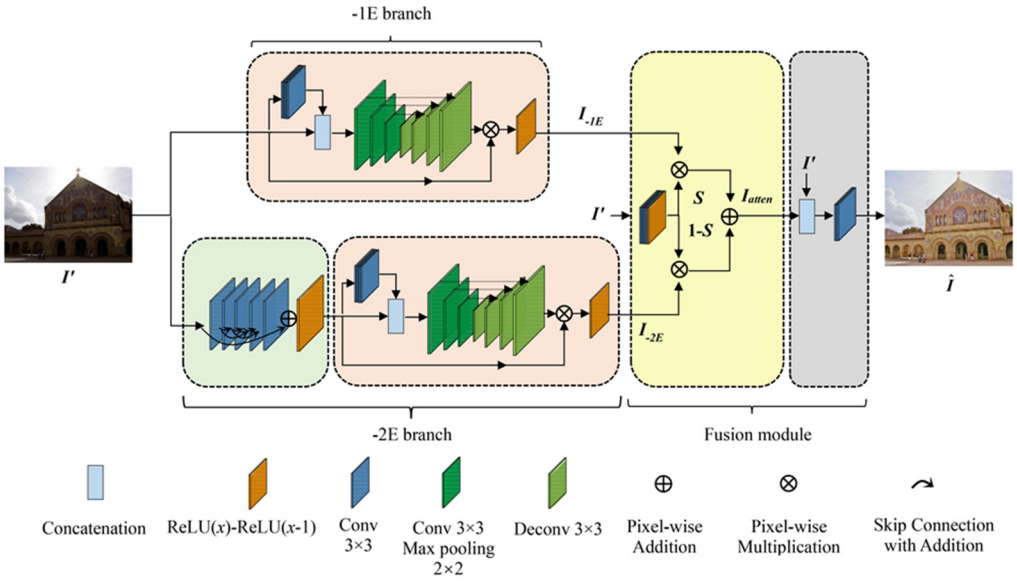


Figure 4. The network architecture of a low-light image enhancement model [35].

Basic enhancement module. F_{en}^{branch} alone constructs the -1E branch without an extra denoising method, and the main form of the -2E branch. The result of the enhancement module is represented as:

$$\boldsymbol{I}_{out}^{branch} = \boldsymbol{I}_{in}^{branch} \circ F_{en}^{branch}\left(\boldsymbol{I}_{in}^{branch}\right) \tag{1}$$

where branch $\in$ {-1E, -2E}. $\boldsymbol{I}_{in}$ and $\boldsymbol{I}_{out}$ are the input and output images, respectively. First, four convolutional layers are utilized for the input image to obtain its additional features, which are subsequently concatenated with the input low-light images before being fed into this enhancement module [35].

Preprocessing module. This module is trained in the -2E branch to separate lightly and heavily degraded images, including natural noise as the primary culprit. The preprocessing module is expressed by applying multilayer element-wise summations. Five convolutional layers with a filter size of 3×3 were applied, and their feature maps were combined with those of the previous layers to assist in the training process. Further, no activation function was implemented after the convolution layer, and only the modified *ReLU* function in the last layer was used to decrease the input properties to the range [0, 1].

$$Out(x) = ReLU(x) - ReLU(x - 1) \tag{2}$$

The range of the estimated noise was set as $(-\infty, +\infty)$ to reproduce the complex designs under low-light conditions.

Fusion module. In this module, the results enhanced by the two-branch network are first merged in the attention unit and subsequently cleaned in the refining unit to produce the final result. Four convolutional layers were applied in the attention unit to generate the attention map $\boldsymbol{S} = F_{atten}(\boldsymbol{I'})$ on the -1E enhanced image, and the equivalent element $1 - S$ for the -2E image, where $\boldsymbol{S}(x, y) \in [0, 1]$. This method aims to continuously assist in the construction of a self-adaptive fusion procedure by modifying the weighted template. The R, G, and B color channels received equal weights provided by the attention map. The results of the attention unit $\boldsymbol{I}_{atten}$ were calculated as follows:

$$\boldsymbol{I}_{atten} = \boldsymbol{I}_{\text{-1E}} \circ \boldsymbol{S} + \boldsymbol{I}_{\text{-2E}} \circ (1 - \boldsymbol{S}) \tag{3}$$

However, the disadvantage of this simple technique is that there may be a loss of certain essential features during the fusion process because the enhanced images from the -1E and -2E branches are generated independently. In addition, owing to the use of a direct metric, there may be an increase in noise. Thus, to address this, $\boldsymbol{I}_{atten}$ is sent to the refining unit $\boldsymbol{F}_{ref}$ with its low-light input concatenated. Finally, the enhanced image is formulated as:

$$\hat{\mathbf{I}} = \boldsymbol{F}_{ref}\left(concat\{\boldsymbol{I}_{atten}, \boldsymbol{I}'\}\right) \tag{4}$$

Loss Function. The combination of three loss functions such as SSIM, VGG, and Smooth was used. SSIM loss estimates the contrast, luminance, and structural diversity jointly; it is more relevant as the loss function here compared with the $L1$ and $L2$. The SSIM loss function is expressed as follows:

$$\mathcal{L}_{SSIM} = 1 - SSIM\left(\hat{\mathrm{I}},\ \mathrm{I}\right) \tag{5}$$

VGG loss is used for addressing two problems. First, when two pixels are constrained with pixel-level distance, one pixel may take the value of any pixels inside the error radius, meaning that this restriction is actually tolerant of possible shifts in the colors and color depth as stated in [35]. Second, since the ground truth is obtained using a mixture of various off-the-shelf enhancement methods, pixel-level loss functions cannot represent the desired quality correctly. It can be formulated as:

$$\mathcal{L}_{VGG} = \frac{1}{WHC}\left\|\mathcal{F}_{VGG}\left(\hat{\mathrm{I}}\right) - \mathcal{F}_{VGG}(\mathrm{I})\right\|^2 \tag{6}$$

where W, H, and C indicate the three dimensions of an image, respectively. The mean squared error was utilized to measure the distance between these features.

Smooth loss can also use total variation loss to describe both the structural features and the smoothness of the estimated transfer function, which is

$$\mathcal{L}_{SMOOTH} = \sum_{branch\{-1E,-2E\}} \left\|\nabla_{x,y}\mathcal{F}_{en}^{branch}(\mathrm{I}_{in})\right\| \tag{7}$$

where $\nabla_{x,y}$ denotes horizontal and vertical per-pixel difference. The combination of these above three loss functions are expressed as:

$$\mathcal{L} = \mathcal{L}_{SSIM} + \lambda_{vl}\cdot\mathcal{L}_{VGG} + \lambda_{sl}\cdot\mathcal{L}_{SMOOTH} \tag{8}$$

Training Data. The low-light image enhancement model was trained using Cai et al. [64] and Low-Light (LOL) datasets [65]. The value of the λ_{vl} was set to zero during the training of the -1E and -2E branches and increased to 0.1 in the joint training stage while λ_{sl} was set to 0.1 as a constant during all training. All Cai and LOL datasets were divided into training set and evaluation set. Cai dataset's images were scaled to one-fifth of the original size and then 10 patches of 256 $\times$ 256 were randomly cropped for the underexposure images of each scene. LOL dataset's images were cropped three patches for each of the images. Finally, the experiments were carried out with combination of 14,531 patches from the Cai dataset and 1449 patches from the LOL dataset.

Figure 5 shows an example of a low-light image enhancement model. The results obtained from the low-light image enhancement model were further fed into the object detection and recognition model.

Figure 5. The results of a low-light image enhancement model using a LOL dataset [65].

3.2. Object Detection and Recognition Model

To realize the object and recognition, a transformer-based encoder–decoder design [36], which is a popular design for sequence prediction, was applied. The self-attention approaches of transformers, which accurately model the interactions of elements in a sequence, render these designs particularly appropriate for collection prediction constraints, such as eliminating duplicate predictions. The Detection Transformer (DETR) predicts all objects at once and is trained end-to-end with a set loss function that achieves bipartite matching between predicted and ground-truth objects [36]. The main difference from several existing detection techniques is that DETR eliminates the need for any customized layers and thus can be regenerated simply in any structure that includes regular CNN and transformer properties. The experimental results showed that DETR achieved more reliable results for detecting large objects. However, in the case of small objects, the detection rate was lower. The network structure of the DETR is simple and is represented in Figure 6. It includes four main parts: (1) a CNN backbone to obtain a short feature description, (2) a transformer encoder, (3) a transformer decoder, and (4) a simple feedforward network (FFN) that produces the last detection prediction.

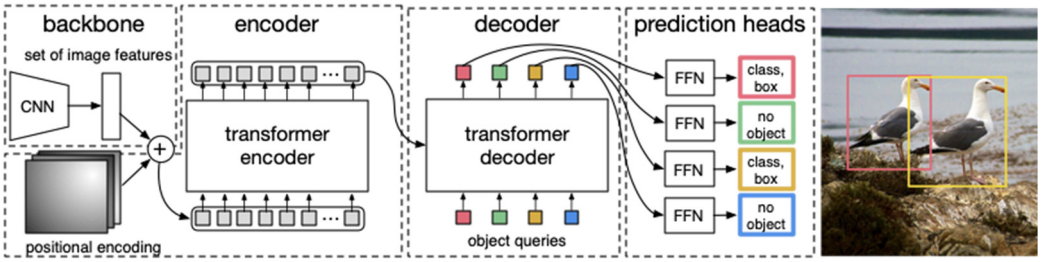


Figure 6. The network structure of the DETR [36].

Backbone. A conventional CNN backbone (ImageNet pretrained ResNet-101) produces a lower-resolution activation map $f \in \boldsymbol{R}^{C \times H \times W}$ from the input image, $x_{img} \in \boldsymbol{R}^{3 \times H_0 \times W_0}$ (with R, G, and B color channels). It is flattened and extended by the model with positional encoding before sending it into a transformer encoder.

Transformer encoder. In this section, first, the channel dimension C of the high-level activation map f is decreased to a small dimension d through a 1×1 convolution filter, and a new $z_0 \in \boldsymbol{R}^{d \times H \times W}$ feature map is created. The transformer encoder waits for the sequence as an input; therefore, the spatial dimensions of z_0 are converted to one dimension, resulting in the creation of a $d \times H \times W$ feature map. Further, each transformer encoder layer has a standard architecture and includes a multihead self-attention module and an FFN.

Transformer decoder. The decoder follows the standard structure of the transformer, converting N embeddings of size d by applying multiheaded self-attention and encoder–

decoder attention mechanisms. However, the N input embeddings must be different to create different results because the decoder is permutation-invariant. These input embeddings are determined positional encodings known as object queries, and they are added to the input of each attention layer in a manner similar to that as the encoder. Subsequently, the decoder transforms N object queries into output embedding. Thereafter, they are independently decoded via an FFN into box coordinates and class labels, producing N final predictions. The model analyzes all objects using pair-wise relationships between them by applying self-attention and encoder–decoder attention over these embeddings [36].

Prediction of Feed-Forward Networks. A three-layer perceptron with a *ReLU* activation function and hidden dimension d, as well as a linear projection layer, computes the final prediction. The normalized center coordinates, height, and width of the box with respect to the input image are predicted using the FFN, whereas the linear layer applies a *softmax* function to predict the class label. Owing to the prediction of a fixed-size set of N bounding boxes, where N is typically much larger than the actual number of objects of interest in an image, an additional special class label *NO* is utilized to indicate that no object is detected within a slot [36].

Loss Function. For auxiliary decoding losses it is convenient to use auxiliary losses [66] in the decoder during training, especially to assist the model in making the correct number of objects of each class. Prediction FFNs and Hungarian loss are added after each decoder layer.

Training Data. For training and evaluation COCO 2017 detection and panoptic segmentation datasets [67,68] are used. These datasets include 118k training images and 5k validation images. Bounding boxes and panoptic segmentation are used to label each picture. In the training set, there is an average of seven instances per image, with up to 63 occurrences in a single image, ranging in size from tiny to huge.

We experimented with an object detection and recognition model on the challenging ExDark [69] dataset. Figure 7 shows the experimental results. Subsequently, the output of the object detection and recognition model is further sent to the TTS model to generate voice feedback for blind users.

Figure 7. The results of the object detection and recognition model on the ExDark [69] dataset.

3.3. Salient Object Detection Model

We followed a two-level nested U-structure network for salient object detection [37]. Qin et al. proposed a residual U-block that includes ReSidual U-block (RSU) which has three primary components as illustrated in Figure 8: (1) an input convolution layer that converts the input feature map $x(H \times W \times C_{in})$ to an intermediate map $F_1(x)$ with a C_{out} channel, used for local feature extraction; (2) a U-Net-like symmetric encoder–decoder architecture with a height of seven that learns to extract and encode the multiscale contextual information $U(F_1(x))$ from the intermediate feature map $F_1(x)$, and (3) a residual connection that combines local features and the multiscale features via the summation $F_1(x) + U(F_1(x))$. The formula in the residual block can be summarized as $H(x) = F_2\ (F_1(x))$

+ x, where $H(x)$ indicates the desired mapping of the input features x; F_2, F_1 stand for the weight layers, which are convolution operations in this setting.

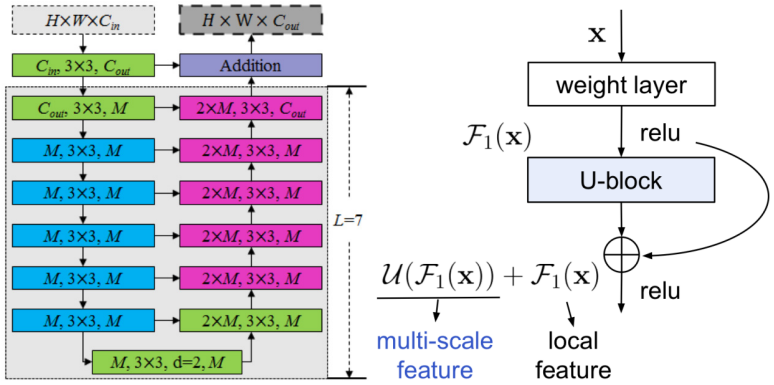


Figure 8. The structure and detail formulation of ReSidual U-block [37]. Larger L leads to deeper residual U-block.

To avoid the disadvantages of CNN network architecture with many nested, such as high computation and complexity to be employed in a real application, the two-level nested U-structure network comprised 11 stages, with each filled by a well-configured residual U-block. Further, the two-level nested U2-Net consisted of three parts: (1) a six-stage encoder, (2) a five-stage decoder, and (3) a saliency map fusion module connected to the decoder stages and the final encoder stage. The design of U2-Net was such that it supports a deep structure with rich multiscale features and has comparatively low memory costs and computation as shown in Figure 9. In encoder stages En_1, En_2, En_3, and En_4, we use residual U-blocks RSU-7, RSU-6, RSU-5, and RSU-4, respectively. As mentioned before, "7", "6", "5", and "4" denote the heights (L) of RSU blocks.

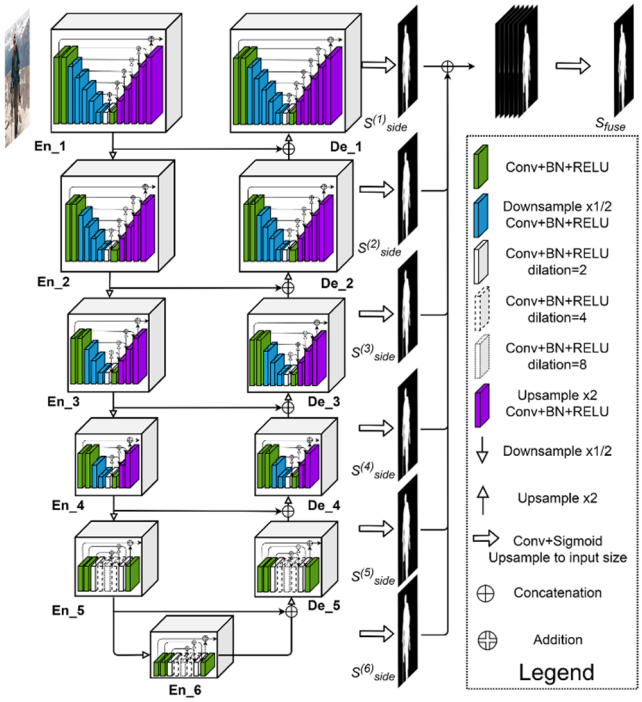


Figure 9. The network architecture of U2-Net model [37].

The decoder stages have similar arrangements to their symmetrical encoder stages concerning En_6. In De_5, the dilated version residual U-block RSU-4F was used. It is similar to encoder stages En_5 and En_6. As input, each decoder stage concatenates the

up-sampled feature maps from the previous stage with those from the symmetrical encoder stage. The saliency map fusion module, which generates saliency probability maps, is the last stage.

Furthermore, the U2-Net architecture is adaptable to a variety of working environments with minimal performance loss because it is based entirely on residual U-blocks with no reliance on any pretrained backbones adapted from image classification. The U2-Net model has versions for computers and embedded devices with sizes of 176.3 and 4.7 MB, respectively.

Training Data. For training and testing, a DUTS-TR dataset—which is a part of DUTS dataset [70]—was used. It is the most-used training dataset for salient object detection and consists of 10,553 images. To make more training images, this dataset was augmented by horizontal flipping and obtained 21,106 images.

After extracting a salient object, we can use a binary mask to obtain the contour of the salient object. These contours are used to provide visually impaired people with visual information in the form of tactile graphics. In certain situations, blind people may not be confident about objects by simply touching their contours. Therefore, we added a method to detect the inner edges of an object from images to aid in better recognition. It is necessary for a blind person to sufficiently recognize a salient object in an image and thus, we applied a binary mask to achieve the internal edges of a salient object using our previous work [38]. First, we perform a salient object by applying its binary mask by creating a matrix with a size and type similar to those of the input image to obtain the desired output image. Subsequently, we copied the non-zero pixels of the binary mask that represent the pixel of the original input image matrix as follows:

$$S_0 = B_m(x, y) * I_i(x, y) \tag{9}$$

where S_0 is the salient object, $B_m(x, y)$ is the binary mask, and $I_i(x, y)$ is the input image. Consequently, we obtained a full-color space-salient object. An example of the masking method is shown in Figure 10. Finally, we could generate the contour and inner edges of a salient object with the added helpful visual information to aid blind people in recognizing the content of an image.

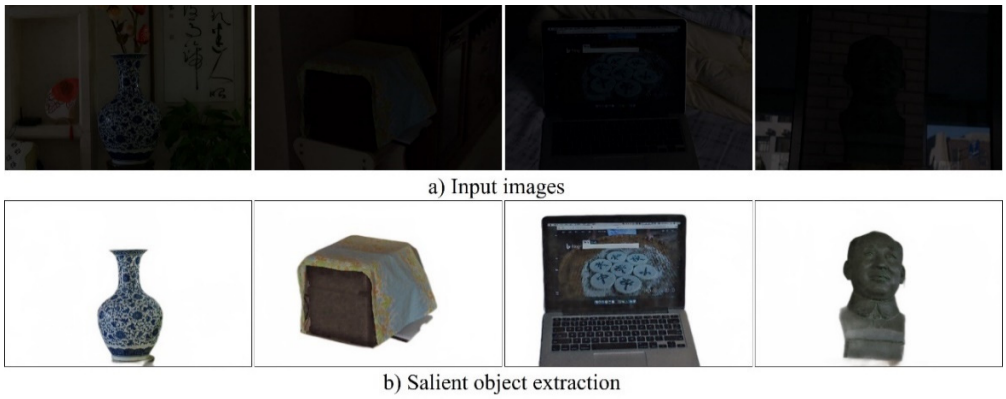


Figure 10. The results of the salient object detection model and binary masking on LOL dataset [65].

3.4. TTS and Tactile Graphics Generation Model

Blind people can receive voice feedback not only regarding surrounding objects, but also about the text data in the natural scene, which are important in our daily lives because they provide the most accurate and unambiguous descriptions of our surroundings, and can also assist blind and visually impaired people in accessing visual information. Text appears on various types of objects in natural scenes, such as billboards, road signs, and product packaging. Scene text contains valuable and high-level semantic information that is required for image comprehension; recognition can be a challenge because of variations in illumination, blurring, color differences, complex backgrounds, poor lighting

conditions, noise, and discontinuity. We used our previous real-time end-to-end scene text recognition [71] as shown in Figure 11 and Tesseract OCR engine [72] to achieve robust and accurate results on ExDark, LOL datasets, and our captured natural scene images. The fundamental part of a text detection and recognition model is a neural network model, which is trained to immediately predict the presence of text occurrences and their geometries from input images. The model is a fully convolutional network modified for text detection that results in dense per-pixel predictions of sentences or text lines. The design can be broken into three parts [71]: the feature extractor, feature merging, and the output layer. The feature extractor can be a convolutional network pretrained on the ImageNet dataset, along with interleaving convolution and pooling layers.

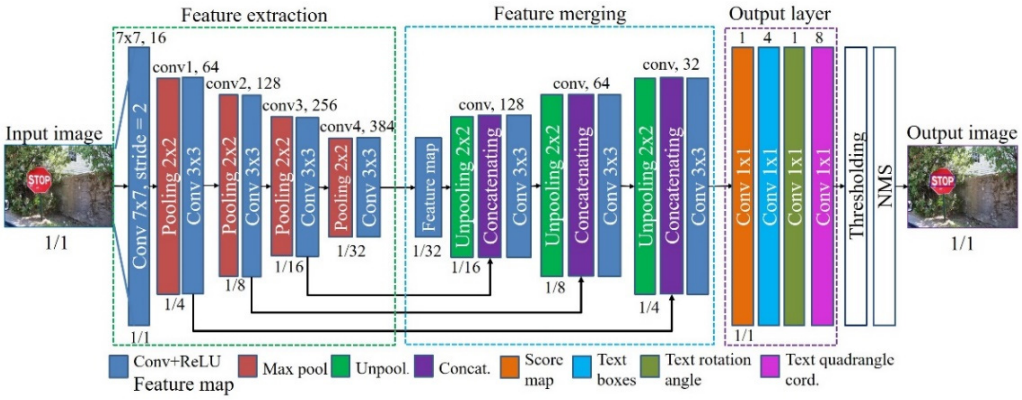


Figure 11. The network architecture of the scene text detection [71].

Texts were recognized by the trained Tesseract OCR model and sent to a TTS for pronunciation. Figure 12 shows an example of the text detection and recognition methods.

Figure 12. The results of the text detection and recognition for TTS on ExDark dataset.

Another difference between our smart glass system and other existing systems is the added function of creating tactile graphics, which provides the blind with visual information regarding the contours of salient objects. As shown in Figure 13, we created tactile graphics of salient objects using our previous work [73] and employed the tactile display system software [63] to assist the blind and visually impaired in perceiving and recognizing natural scene images.

Figure 13. The results of the tactile graphics generation on LOL dataset.

A refreshable 2D multiarray Braille display was used to dynamically represent the tactile graphics of salient objects. The tactile display has 12 × 12 Braille cells, and its simulator is illustrated in Figure 14. Further, the volume control buttons are located on the left side and can be used to adjust the volume of audio or TTS and the speed of the TTS can be increased or decreased with a long click. In addition, other buttons to control various tasks are included, as shown in Figure 14.

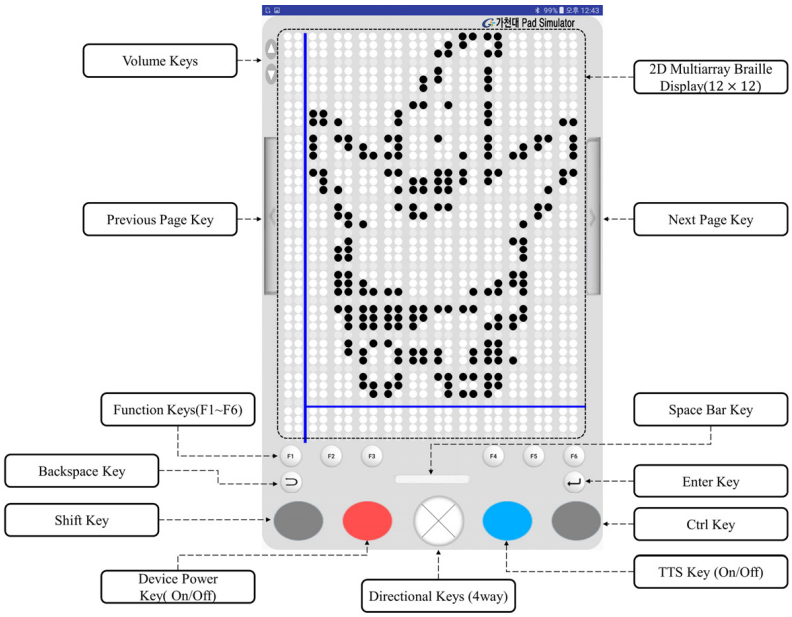


Figure 14. The design of tactile display simulator [63].

4. Experiments and Results

In this section, we present the results of the models on the artificial intelligence server. Experimental validations of the proposed smart glass system were conducted in a night-time environment, and object detection, salient object extraction, text recognition, and tactile graphics generation were focused upon. The challenging LOL dataset [65] comprising 500 low-light images and the ExDark dataset [69] comprising 7363 night images were employed. As embedded systems may not be the optimal option to increase the energy storage viability of smart glasses and ensure the real-time performance of the system, using a high-performance artificial intelligence server is more effective [74].

The performance of the artificial intelligence server determines whether the proposed smart glass system succeeds or fails. This is because the deep learning models employed in smart glass systems consume a significant amount of computing resources on an artificial intelligence server. Thus, to evaluate the performance of the proposed smart glass system, we conducted experiments using an artificial intelligence server, and the system environment is shown in Table 2.

Table 2. The environment of the artificial intelligence server.

Item	Specifications	Details
GPU	GPU 2-GeForce RTX 2080 Ti 11 GB	Two GPU are installed
CPU	Intel Core 9 Gen i7-9700k (4.90 GHz)	
RAM	DDR4 64 GB (DDR4 16GB × 4)	Samsung DDR4 16 GB PC4-21300
Storage	SSD: 512 GB/HDD: TB (2 TB × 2)	
Motherboard	ASUS PRIME Z390-A STCOM	
OS	Ubuntu Desktop	version: 18.0.4 LTS
LAN	port 1 (internal)—10/100 Mbps port 2 (external)—10/100 Mbps	
Power	1000 W (+12 V Single Rail)	Micronics Perform. II HV 1000 W Bronze

The artificial intelligence server received captured images from a local part consisting of a smartphone and smart glass. Thereafter, the received images were processed using computer vision and deep learning models. The final results were sent to the local part through Wi-Fi/Internet connection, and the user could hear the output audio information via a speaker or earphone or perceive tactile graphics using the refreshable tactile device. The experimental results of the deep learning models running on the artificial intelligence server have been presented below.

4.1. Experimental Results of Object Detection Model

First, we evaluated the performance of the object detection model, which is one of the most essential aspects of the proposed system. The object detection model was trained with AdamW [75], with initial transformer's learning rate to 10^{-4}, the backbone's to 10^{-5}, and weight decay to 10^{-4}. Before experimenting on LOL dataset, we obtained the results on COCO 2017 dataset with two varying backbones: a ResNet-50 and a ResNet-101 and compared with Faster R-CNN [76] model. The corresponding models are called, respectively, DETR-R50 and DETR-R101. In this comparison, we used an average precision (AP) metric as explained in [77]. Following [36], we also increased the feature resolution by adding a dilation to the last stage of the backbone and removing a stride from the first convolution of this stage. The corresponding models are called, respectively, DETR-DC5-R50 and DETR-DC5-R101 (dilated C5 stage). Table 3 shows a full comparison of floating point operations per second (FLOPS), frame per second (FPS), average precision (AP) of object detection with transformers (DETR), and Faster R-CNN as explained in [36].

Blind people desire to learn about the world around them during their travel, whether during daytime or night-time. Till now, object detection approaches have been efficient in environments with sufficient illumination; however, low light and a lack of illumination are among the main problems of object detection models. To address this issue, we used the low-light enhancement approach and subsequently detected objects to assist the blind user in traveling independently at any time of the day.

Table 3. The performance comparison of DETR with Faster R-CNN with ResNet-50 and ResNet-101 backbones on the COCO 2017 validation set. The results of Faster R-CNN models in Detectron2 [78] and GIoU [79] are shown in the top three rows and middle three rows, respectively. DETR models achieve comparable results to heavily tuned Faster R-CNN baselines, having lower AP_S but greatly improved AP_L. S: small objects, M: medium objects, L: large objects.

Models	GFLOPS/FPS	#Params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster RCNN-R50-DC5	320/16	166M	39.0	60.5	42.3	21.4	43.5	52.5
Faster RCNN- R50-FPN	180/26	42M	40.2	61.0	43.8	24.2	43.5	52.0
Faster RCNN-R101-FPN	246/20	60M	42.0	62.5	45.9	25.2	45.6	54.6
Faster RCNN- R50-DC5+	320/16	166M	41.1	61.4	44.3	22.9	45.9	55.0
Faster RCNN- R50-FPN+	180/26	42M	42.0	62.1	45.5	26.6	45.4	53.4
Faster RCNN-R101-FPN+	246/20	60M	44.0	63.9	**47.8**	**27.2**	48.1	56.0
DETR-R50	86/28	41M	42.0	62.4	44.2	20.5	45.8	61.1
DETR-DC5-R50	187/12	41M	43.3	63.1	45.9	22.5	47.3	61.1
DETR-R101	152/20	60M	43.5	63.8	46.4	21.9	48.0	61.8
DETR- DC5-R101	253/10	60M	**44.9**	**64.7**	47.7	23.7	**49.5**	**62.3**

We evaluated the performance of the object detection models on a low-light image following the application of the low-light enhancement method. We compared the DETR model with other 10 state-of-the-art models such as OHEM [80], Faster RCNNwFPN [81], RetinaNet [82], RefineDet512+ [83], RFBNet512-E [84], CornerNet511 [85], M2Det800 [86], R-DAD-v2 [87], ExtremeNet [88], and CenterNet511 [89]. We used the results in their papers and their source code for performance comparison. We performed quantitative analysis by using metrics such as Precision, and Recall, as in our earlier studies [38,71,90] and AP. Precision and recall rates could be obtained by comparing pixel-level ground truth images with the results of the proposed method and calculated as follows:

$$Precision_{C_{ij}} = \frac{TP_{C_{ij}}}{TP_{C_{ij}} + FP_{C_{ij}}} \tag{10}$$

$$Recall_{C_{ij}} = \frac{TP_{C_{ij}}}{TP_{C_{ij}} + FN_{C_{ij}}} \tag{11}$$

where $Precision_{C_{ij}}$ represents the Precision of category *Ci* in the *j* th image, while $Recall_{C_{ij}}$ represents the Recall of category *Ci* in the *j* th image, *TP* denotes the number of true positives indicating correctly detected object regions, *FP* denotes the number of false positives, and *FN* denotes the number of false negatives. *Precision* is defined as the number of true-positive pixels over the number of true-positive pixels plus the number of false-positive pixels. Recall is defined as the number of true-positive pixels over the number of true-positive pixels plus the number of false-negative pixels. The Average Precision (AP) of the category *Ci* can be calculated as follows:

$$AP_{C_{ij}} = \frac{1}{m} \sum_{j=1}^{m} Precision_{C_{ij}} \tag{12}$$

The comparison results of the DETR and other state-of-the-art models which are published top conferences and journals including CVVR, ICCV, ECCV, and AAAI in the recent years are presented in Table 4. As we can see, object detection with Transformers achieves the best performance on datasets LOL and ExDart in terms of AP_{50}, AP_{75}, AP_M, and AP_L evaluation metrics. DETR achieves the second-best overall performance which is slightly inferior to CenterNet511 and M2Det800 in terms of only AP and APS evaluation metrics, respectively.

Table 4. The performance comparison of DETR with other state-of-the-art methods on LOL and ExDark datasets. The best results are marked with bold. S: small objects, M: medium objects, L: large objects.

Models	Backbone Network	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
OHEM [80]	VGGNet16	21.3	40.3	21	4.3	21.5	36.8
Faster RCNN w FPN [81]	ResNet101-FPN	35.4	57.9	38.3	16.4	37.4	47.3
RetinaNet [82]	ResNeXt101-FPN	38.6	59.3	43.2	22.8	43.4	50.5
RefineDet512+ [83]	ResNet101	40.1	61.3	43.8	24.4	43.8	52.9
RFBNet512-E [84]	VGGNet16	33.1	53.2	34.7	15.9	36.1	46.3
CornerNet511 [85]	Hourglass104	40.8	55.7	43.4	19.6	43.2	55.7
M2Det800 [86]	VGGNet16	42.3	62.8	47.6	**27.5**	46.3	53.8
R-DAD-v2 [87]	ResNet101	42.7	61.6	46.8	23.5	43.6	53.2
ExtremeNet [88]	Hourglass104	41.5	59.2	46.3	22.8	45.2	55.9
CenterNet511 [89]	Hourglass104	**46**	63.1	48.5	26.8	48.3	57.2
DETR [36]	ResNet101	45.3	**63.5**	**50.3**	26.4	**48.9**	**60.7**

Figure 15 shows the results of the object detection model on the challenging LOL dataset. The experimental results indicated that in low-light images, the object detection model could correctly detect certain objects, while a few were detected incorrectly or could not be detected at all. However, more objects were correctly detected following the image illumination enhancement. The first row presents low-light images such as people, chairs, TVs, books, and different types of objects. The second and third rows display the results of the object detection model before and after the application of the low-light enhancement method, respectively.

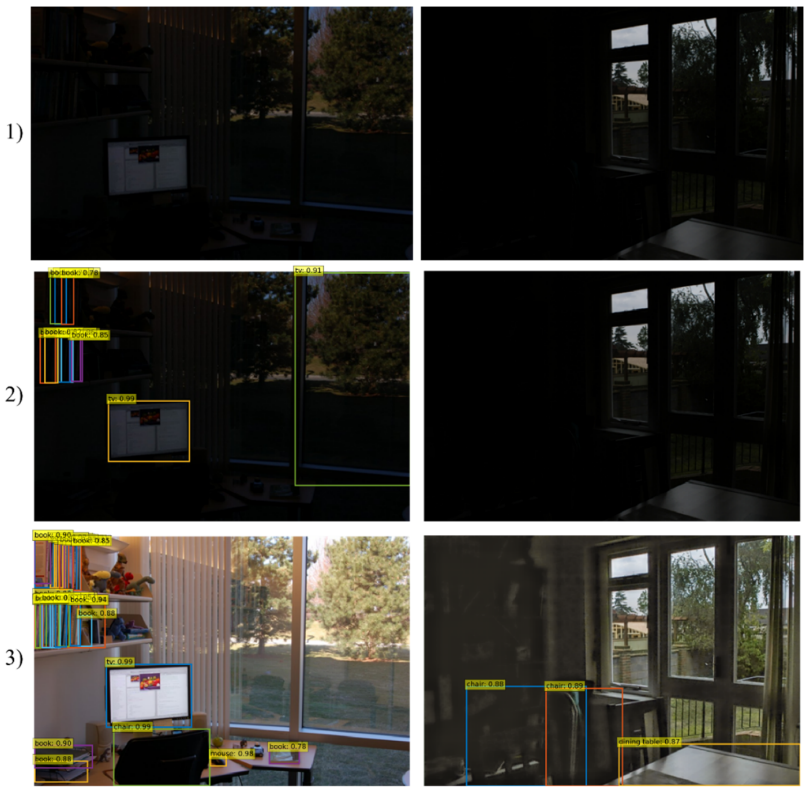


Figure 15. The results of object detection model on the challenging LOL datasets. (**1**) low-light input images; (**2**) object detection before image enhancement; (**3**) object detection after image enhancement.

Thus, the experimental results show that the object detection model performed well and accurately after image enhancement. Furthermore, it worked effectively, even when multiple objects were present, as shown in Figure 15. The data of the recognized objects were converted to audio and sent to the local part via the network.

4.2. Experimental Results of Salient Object Extraction Model

Second, we experimentally evaluated the performance of a salient object extraction model, which is one of the most significant steps in the process of creating tactile graphics from natural scene images for BVI people. Although the effective aspects and applications of salient object extraction have been emphasized by many researchers, the detection of salient objects from dark light images has not been sufficiently studied. We employed low-light image enhancement and salient object extraction models to create simple and easy-to-understand tactile graphics from low-light and dark images. As a result, BVI people could hear the name of the object around them and feel its contour via a refreshable tactile display.

To comprehensively evaluate the quality of salient object extraction methods, we additionally calculated the F-measure (FM) value, which balanced measurements between the mean of precision and recall rates and maximal F-measure (*maxFM*), weighted F-measure (*WFM*), and mean absolute error (*MAE*) metric as explained in [77]. A higher F-measure meant a higher performance and this was expressed as follows:

$$FA = \frac{(1 + 0.3) \times Precision \times Recall}{0.3 \times Precision + Recall} \tag{13}$$

A perfect match occurs when F-measure = 1 and the closer to 1 the F-measure gets, the better the detection is considered. *MAE* denotes the average per-pixel difference between a predicted saliency map and its ground truth mask. It is defined as:

$$MAE = \frac{1}{H \times W} \sum_{r=1}^{H} \sum_{c=1}^{W} |PM(r, c) - GT(r, c)| \tag{14}$$

where *PM* and *GT* are the probability map of the salient object detection and the corresponding ground truth, respectively; (*H*, *W*) and (*r*; *c*) are the (height, width) and the pixel coordinates. *WFM* is applied as a complementary measure to *maxFM* for overcoming the possible unfair comparison caused by "interpolation flaw, dependency flaw and equal-importance flaw". It is formulated as:

$$WFM = (1 + 0.3) \frac{Precision^w \times Recall^w}{0.3 \times Precision^w + Recall^w} \tag{15}$$

Table 5 shows the comparison results of three evaluation metrics and state-of-the-art performance of 10 various models which were published in top conferences such as CVVR, ICCV, and ECCV. As we can see, U2-Net obtained the best results on datasets LoL and ExDark in terms of all of the three evaluation metrics.

Further, similar to the object detection model above, the salient object extraction model first with a low-light image and subsequently after applying the low-light enhancement method were visually compared. In Figure 16, the first row shows the dark images considered, such as a flowerpot, clothes, and a microwave oven. The second row displays the salient object extraction before the low-light enhancement method. Further, the third and fourth rows show the results of the salient object extraction after the image enhancement method and salient objects in full-color space using the binary masking technique, respectively. As shown in the second row of Figure 16, the salient object extraction results from dark images exhibit incorrect extraction owing to the similar background and foreground. In contrast, the proposed salient object extraction method can reduce these drawbacks. With the help of the low-light image enhancement method, we increased the difference between the background and the object and thus efficiently extracted multiple objects.

Moreover, enhancing low-light image illumination also increases the accuracy of detecting the inner edges of salient objects using the edge detection method.

Table 5. The performance comparison of U2-Net with other state-of-the-art models on LOL and ExDark datasets. The best results are marked in bold.

Models	Backbone Network	maxFM	MAE	WFM	Published
Amulet [91]	VGGNet16	0.736	0.103	0.624	ICCV17
RAS [92]	VGGNet16	0.748	0.094	0.683	ECCV18
PiCANet [93]	VGGNet16	0.763	0.079	0.675	CVPR18
AFNet [94]	VGGNet16	0.772	0.064	0.687	CVPR19
MSWS [95]	Dense-169	0.685	0.108	0.614	CVPR19
SRM [96]	ResNet50	0.759	0.081	0.629	ICCV17
PiCANetR [93]	ResNet50	0.786	0.073	0.647	CVPR18
CPD [97]	ResNet50	0.765	0.062	0.689	CVPR19
PoolNet [98]	ResNet50	0.784	0.065	0.691	CVPR19
BASNet [99]	ResNet34	0.792	0.061	0.706	CVPR19
U2-Net [37]	RSU	**0.814**	**0.058**	**0.725**	CVPR20

Figure 16. The results of salient object extraction model on the challenging LOL datasets. (**1**) low-light input images; (**2**) salient object extraction before image enhancement; (**3**) salient object extraction after image enhancement; (**4**) salient objects in full color space.

It is essential for BVI people to fully perceive a salient object with outer and inner edges in a natural scene. Therefore, we used our previous work [38] to obtain the salient objects in a full-color space and further inner edge detection.

4.3. Experimental Results of Text-to-Speech Model

Finally, we experimentally evaluated the performance of the text-to-speech model. Text data are now encountered in all aspects of our daily lives. Therefore, conveying the text information to BVI people through audio to detect objects and convey their contours through tactile graphics is crucial. Based on these models, the BVI users can hear visual information from the natural scene around them, as shown in Figure 17.

In this study, we focused on text recognition from natural scene images in a dark environment. Because text recognition from a document or scanned images, paper documents, and books have achieved remarkable results, we used the ExDark dataset to evaluate the experimental results. We used Precision, Recall, and F-measure evaluation metrics to compare text detection and recognition models. The text detection results of our previous method and eight other cutting-edge models which were published in top conferences such as CVVR, ECCV, and AAAI are compared in Table 6.

Table 6. The performance comparison of our previous text detection model with other state-of-the-art models on ExDark datasets. The best results are marked in bold.

Methods	Backbone Network	Precision	Recall	FM	Published
Zhang et al. [100]	VGGNet16	0.705	0.414	0.527	CVPR16
Holistic [101]	VGGNet16	0.716	0.563	0.629	CVPR16
SegLink [102]	VGGNet16	0.724	0.578	0.631	CVPR17
He et al. [103]	VGGNet16	0.793	0.784	0.789	CVPR17
EAST [104]	VGGNet16	0.817	0.762	0.796	CVPR17
TextSnake [105]	VGGNet16	0.822	0.786	0.807	ECCV18
PixelLink [106]	VGGNet16	0.837	0.814	0.828	AAAI18
Wang et al. [107]	ResNet50	0.849	0.724	0.785	CVPR18
Our Previous model [71]	VGGNet16	**0.863**	**0.817**	**0.824**	IJWMIP20

The evaluation of the end-to-end system is a combination of both detection and recognition. The first predicted text examples are matched with ground truth examples after comparison of the recognized text content. The performance of end-to-end evaluation matching is initially implemented in a process similar to that of text detection. Our previous text recognition model and seven other state-of-the-art models are compared, using the ExDark dataset, in Table 7.

Table 7. The performance comparison of our previous text recognition model with other state-of-the-art models on ExDark datasets. The best results are marked in bold.

Methods	Recognition (%)	Published
Jaderberg et al. [108]	79.6	CVPR14
Shi et al. [109]	86.3	CVPR16
Shi et al. [110]	87.5	IEEE TPAMI16
Lee et al. [111]	88.2	CVPR16
Jaderberg et al. [112]	88.9	CVPR14
Shi et al. [113]	89.4	IEEE TPAMI18
Cheng et al. [114]	91.5	CVPR17
Our previous model [71]	**92.8**	IJWMIP20

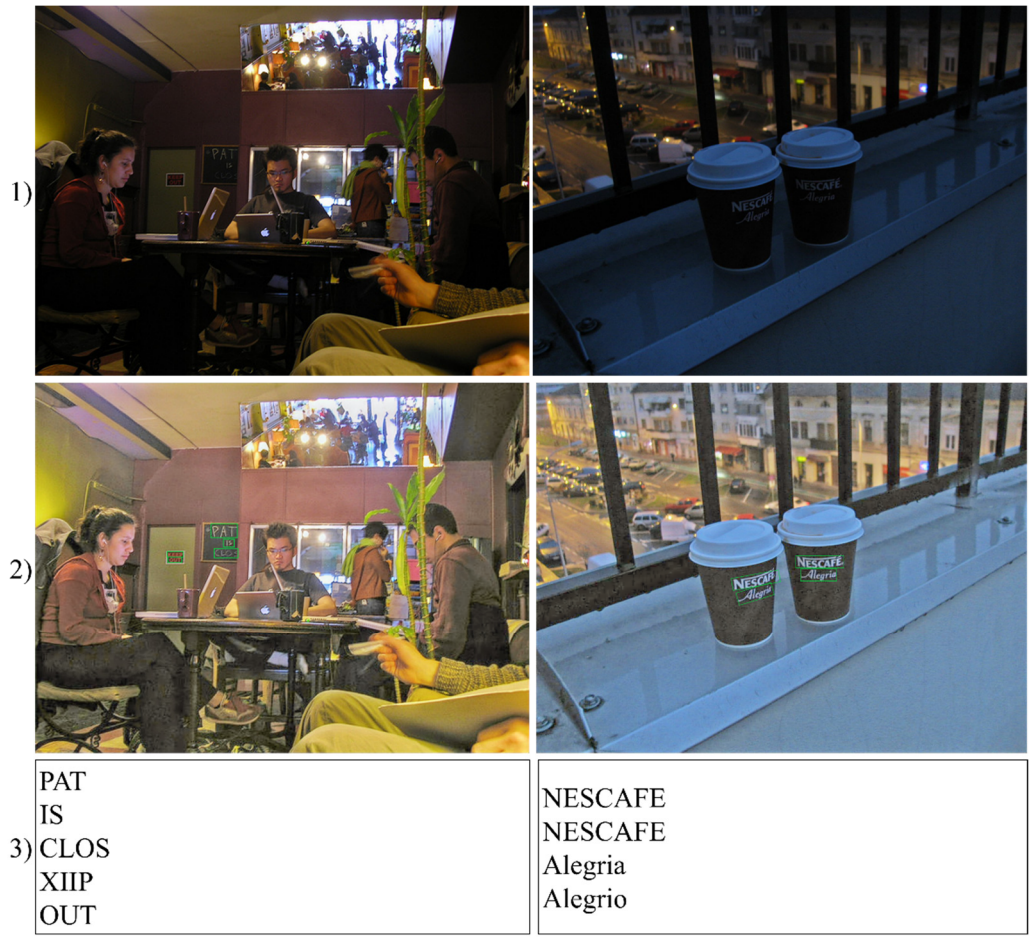


Figure 17. The results of scene text-to-speech model on the challenging ExDark datasets. (**1**) low-light input images; (**2**) text detection; (**3**) recognized text.

Figure 17 shows the results of the scene text-to-speech model obtained for the low-light images. The first row displays input images with dark scenes and different objects such as people, chairs, coffee cups, and teapots. The second and third rows show the results of the text detection method and recognized words respectively. The recognition of certain words had mistakes such as "XIIP" and "Alegrio" because of small character size and the characters being blocked by objects.

To establish the communication between client and server, we utilized gRPC (Google's Remote Procedure Call) protocol. gRPC is a free and open-source protocol that defines the bidirectional communication APIs to organize microservices between client and server. At high level (transport and application), it allows us to specify the format of REQUEST and RESPONSE messages through which the communication will be handled. gRPC protocol is built on top of HTTP/2 and inter-operates with well-known transport protocols such as TCP and UDP. It generates less latency and supports streaming, load balancing, and easy authentication procedures. At the core of gRPC, we need to define the message and services using Protocol Buffers (PB). PB efficiently serializes structured data that we call a payload and is very convenient to transport a lot of data. We also obtained the performance of frame processing time for each stage including Bluetooth image transmission between smart glass and smartphone, 5G/WiFi image transmission time between smartphone and server, and four models' image processing time in the artificial server. Table 8 presents the average processing time in seconds to perform each stage. As we can see, the total time for all stages is 0.936 s which is relevant for real-life situations.

Table 8. The performance of average frame processing time (in seconds) per sequence. The average input image size is 640 × 456.

Image Transmission and Processing	Average Processing Time (sec)
Bluetooth image transmission (between smart glass and smartphone)	0.047
5G/Wi-Fi image transmission (between smartphone and server)	0.024
Low-light image enhancement model	0.051
Object recognition model	0.173
Salient object extraction and tactile graphics model	0.215
Text recognition and TTS model	0.426
Total	**0.936**

We compared the proposed smart glass system with the other similar works in the field of wearable assistive technologies for BVI. The comparison results of the main features of different assistive systems are shown in Table 9.

Table 9. The comparison with cutting edge systems.

Systems	Image Dataset	Working Architecture	Coverage Area	Connection	Components	Results
Daescu et al. [13]	VGGFace2	Client–server	Outdoor and Indoor	5G	Smart glass, Phone, Server	Face recognition
Anandan et al. [16]	No Dataset	Local (Embedded)	Outdoor and Indoor	Offline	Raspberry-Pi, Camera, GPS	Obstacle detection
Joshi et al. [17]	Local Dataset	Local (Embedded)	Outdoor and Indoor	Offline	Distance Sensor, DSP, Camera	Object and text recognition and obstacle detection
Crose et al. [18]	No Dataset	Local (Smartphone)	Outdoor and Indoor	Offline	Smartphone, Pedestrian Dead Reckoning	Navigation
Park et al. [32]	COCO 2017, Local Dataset	Client–server	Outdoor and Indoor	NA	Raspberry-Pi, Camera	Object recognition and obstacle detection
Pardasani et al. [33]	No Dataset	Local (Embedded)	Outdoor and Indoor	Offline	Raspberry-Pi, Camera	Object and text recognition and obstacle detection
Bai et al. [34]	No Dataset	Local (Embedded)	Outdoor and Indoor	Offline	Depth Camera, Smart glass, CPU board	Obstacle detection
Mandal et al. [39]	No Dataset	Local (Google Glass)	Outdoor and Indoor	Offline	Google Glass	Face recognition
Chen et al. [40]	Labeled Faces in the Wild and PASCAL VOC	Client–server	Outdoor and Indoor	4G/Wi-Fi	Raspberry-Pi, Camera	Face, Object and text recognition
Lee et al. [45]	Local dataset	Client–server	Outdoor and Indoor	Wi-Fi	Smart glasses, phone	Face recognition
Yang et al. [115]	ADE20K, PASCAL VOC	Local (Laptop)	Outdoor and Indoor	Offline	Depth Camera, Smart glass, Laptop	Obstacle detection
Mancini et al. [116]	No Dataset	Local(Embedded)	Outdoor	Offline	Camera, PCB, and vibration motor	Obstacle detection
Patil et al. [117]	No Dataset	Local(Embedded)	Outdoor and Indoor	Offline	Sensors, Vibration motors	Obstacle detection
Al-Madani et al. [118]	No Dataset	Local(Embedded)	Indoor	Offline	BLE fingerprint, fuzzy logic	Localization in building
Our System	COCO 2017, LOL, Exdark	Client–server	Outdoor and Indoor (Nigh-time)	5G/Wi-Fi	Smart Glass, Phone, Refreshable Braille display	Object, text recognition, Tactile graphics

In addition, we obtained the experimental results using all models in the smart glass system for the sake of simplicity. The results are shown in Figure 18. The first and second columns show dark input images and the results of the image enhancement technique, respectively. The results of object detection, salient object extraction, and text detection, which are the main models of the proposed system, are shown in the third, fourth, and sixth columns, respectively. Further, the fifth column displays the results of the contour detection method used to create the tactile graphics. In the last column, recognized text from text detection is presented. The images need to be zoomed in on in order to see the specific and detailed results.

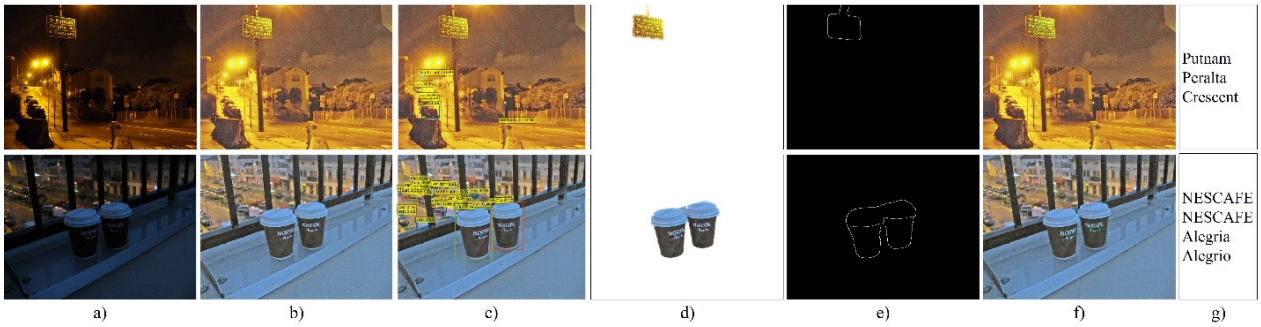


Figure 18. The result of each model of smart glass system. (**a**) Input image; (**b**) low light image enhancement; (**c**) object detection; (**d**) salient object extraction; (**e**) contour detection for tactile graphic; (**f**) text detection; (**g**) recognized text.

5. Limitation and Discussion

In addition to the aforementioned achievements, the proposed system has certain shortcomings. These drawbacks can be found in object detection, salient object extraction, and text recognition models, and experimental results with these drawbacks are shown in Figures 15–17. In certain situations, the object detection model detects more than ten objects, where a few of them are small objects or incorrectly detected, as shown in Figure 15. Further, the salient object extraction model may incorporate certain errors in extracting the regions for the cases where the image pixel values were quite close to each other, as shown in Figure 16. Furthermore, the texts were recognized from natural scene images with certain errors owing to the small size of characters, orientation, and characters being blocked by other objects, as shown in Figure 17.

Furthermore, this study covers only the artificial intelligence server part of the smart glass system and the hardware perspective that is the local part of the system and the experiments with BVI people could not be investigated owing to device patenting, pandemic, and other circumstances. We believe that in the near future, we will find solutions to these problems, conduct experiments in fully integrated software and hardware, and bring convenience to the lives of the BVI.

6. Conclusions

This paper describes a smart glass system that includes object detection, salient object extraction, and text recognition models using computer vision and deep learning for BVI people. The proposed system is fully automatic and runs on an artificial intelligence server. It detects and recognizes objects from low-light and dark-scene images to assist BVI in a night-time environment. The traditional smart glass system was extended using deep learning models and the addition of salient object extraction for tactile graphics and text recognition for text-to-speech.

Smart glass systems require greater energy and memory in embedded systems because they are based on deep learning models. Therefore, we built it in an artificial intelligence server to ensure real-time performance and solve energy problems. With the advancement of the 5G era, transmitting image data to a server or receiving real-time results for users is no longer a concern. The experimental results showed that object detection, salient object

extraction, and text recognition models were robust and performed well with the help of low-light enhancement techniques in a dark scene environment. In the future, we aim to create low-light and dark-image datasets with bounding box and ground truth data to address object detection and text recognition tasks as well as evaluations at night

Author Contributions: Conceptualization, data curation, writing—original draft, data curation, and investigation: M.M.; project administration, supervision, and writing—review and editing: J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2019R1F1A1057757).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Steinmetz, J.D.; Bourne, R.R.; Briant, P.S.; Flaxman, S.R.; Taylor, H.R.; Jonas, J.B.; Abdoli, A.A.; Abrha, W.A.; Abualhasan, A.; Abu-Gharbieh, E.G.; et al. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: The Right to Sight: An analysis for the Global Burden of Disease Study. *Lancet Glob. Health* **2021**, *9*, e144–e160. [CrossRef]
2. Dunai Dunai, L.; Chillarón Pérez, M.; Peris-Fajarnés, G.; Lengua Lengua, I. Euro banknote recognition system for blind people. *Sensors* **2017**, *17*, 184. [CrossRef] [PubMed]
3. Lee, J.; Ahn, J.; Lee, K.Y. Development of a raspberry Pi-based banknote recognition system for the visually impaired. *J. Soc. E-Bus. Stud.* **2018**, *23*, 21–31.
4. Patrycja, B.-A.; Osiński, D.; Wierzchoń, M.; Konieczny, J. Visual Echolocation Concept for the Colorophone Sensory Substitution Device Using Virtual Reality. *Sensors* **2021**, *21*, 237.
5. Chang, W.-J.; Chen, L.-B.; Sie, C.-Y.; Yang, C.-H. An artificial intelligence edge computing-based assistive system for visually impaired pedestrian safety at zebra crossings. *IEEE Trans. Consum. Electron.* **2020**, *67*, 3–11. [CrossRef]
6. Yu, S.; Lee, H.; Kim, J. Street crossing aid using light-weight CNNs for the visually impaired. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
7. Yuksel, B.F.; Fazli, P.; Mathur, U.; Bisht, V.; Kim, S.J.; Lee, J.J.; Jin, S.J.; Siu, Y.-T.; Miele, J.A.; Yoon, I. Human-in-the-Loop Machine Learning to Increase Video Accessibility for Visually Impaired and Blind Users. In Proceedings of the 2020 ACM Designing Interactive Systems Conference, Eindhoven, The Netherlands, 6–10 July 2020.
8. Liu, X.; Carrington, P.; Chen, X.A.; Pavel, A. What Makes Videos Accessible to Blind and Visually Impaired People? In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021.
9. Spagnol, S.; Hoffmann, R.; Martínez, M.H.; Unnthorsson, R. Blind wayfinding with physically-based liquid sounds. *Int. J. Hum.-Comput. Stud.* **2018**, *115*, 9–19. [CrossRef]
10. Skulimowski, P.; Owczarek, M.; Radecki, A.; Bujacz, M.; Rzeszotarski, D.; Strumillo, P. Interactive sonification of U-depth images in a navigation aid for the visually impaired. *J. Multimodal User Interfaces* **2019**, *13*, 219–230. [CrossRef]
11. Zhao, Y.; Wu, S.; Reynolds, L.; Azenkot, S. A face recognition application for people with visual impairments: Understanding use beyond the lab. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018.
12. Sharma, S.; Jain, S. A static hand gesture and face recognition system for blind people. In Proceedings of the 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN) IEEE, Noida, India, 7–8 March 2019.
13. Daescu, O.; Huang, H.; Weinzierl, M. Deep learning based face recognition system with smart glasses. In Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, Rhodes, Greece, 5–7 June 2019.
14. Gurari, D.; Li, Q.; Lin, C.; Zhao, Y.; Guo, A.; Stangl, A.; Bigham, P.J. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
15. Rocha, D.; Carvalho, V.; Gonçalves, J.; Azevedo, F.; Oliveira, E. Development of an automatic combination system of clothing parts for blind people: MyEyes. *Sens. Transducers* **2018**, *219*, 26–33.
16. Anandan, M.; Manikandan, M.; Karthick, T. Advanced Indoor and Outdoor Navigation System for Blind People Using Raspberry-Pi. *J. Internet Technol.* **2020**, *21*, 183–195.
17. Joshi, R.C.; Yadav, S.; Dutta, M.K.; Travieso-Gonzalez, C.M. Efficient Multi-Object Detection and Smart Navigation Using Artificial Intelligence for Visually Impaired People. *Entropy* **2020**, *22*, 941. [CrossRef]
18. Croce, D.; Giarre, L.; Pascucci, F.; Tinnirello, I.; Galioto, G.E.; Garlisi, D.; Valvo, A.L. An indoor and outdoor navigation system for visually impaired people. *IEEE Access* **2019**, *7*, 170406–170418. [CrossRef]
19. eSight. Available online: https://esighteyewear.com/ (accessed on 28 October 2021).
20. NuEyes Pro. Available online: https://www.nueyes.com/ (accessed on 28 October 2021).
21. OrCam My Eye. Available online: https://www.orcam.com/en/myeye2/ (accessed on 28 October 2021).
22. Oxsight. Available online: https://oxsightglobal.com/ (accessed on 28 October 2021).

23. Oton Glass. Available online: https://www.jamesdysonaward.org/en-GB/2016/project/oton-glass/ (accessed on 28 October 2021).
24. AngleEye. Available online: https://www.closingthegap.com/angeleye-series-angleeye-smart-reader-and-angeleye-smart-glasses// (accessed on 28 October 2021).
25. EyeSynth. Available online: https://eyesynth.com/?lang=en/ (accessed on 28 October 2021).
26. Envision. Available online: https://www.letsenvision.com/envision-glasses/ (accessed on 28 October 2021).
27. Hu, M.; Chen, Y.; Zhai, G.; Gao, Z.; Fan, L. An overview of assistive devices for blind and visually impaired people. *Int. J. Robot. Autom.* **2019**, *34*, 580–598. [CrossRef]
28. Manjari, K.; Verma, M.; Singal, G. A survey on assistive technology for visually impaired. *Internet Things* **2020**, *11*, 100188. [CrossRef]
29. Gupta, L.; Varma, N.; Agrawal, S.; Verma, V.; Kalra, N.; Sharma, S. Approaches in Assistive Technology: A Survey on Existing Assistive Wearable Technology for the Visually Impaired. In *Computer Networks, Big Data and IoT*; Springer: Singapore, 2021; pp. 541–556.
30. El-Taher, F.E.Z.; Taha, A.; Courtney, J.; Mckeever, S. A systematic review of urban navigation systems for visually impaired people. *Sensors* **2021**, *21*, 3103. [CrossRef]
31. Son, H.; Krishnagiri, D.; Jeganathan, V.S.; Weiland, J. Crosswalk guidance system for the blind. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020.
32. Park, H.; Ou, S.; Lee, J. Implementation of Multi-Object Recognition System for the Blind. *Intell. Autom. Soft Comput.* **2021**, *29*, 247–258. [CrossRef]
33. Pardasani, A.; Prithviraj, N.I.; Banerjee, S.; Kamal, A.; Garg, V. Smart assistive navigation devices for visually impaired people. In Proceedings of the 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 23–25 February 2019.
34. Jinqiang, B.; Lian, S.; Liu, Z.; Wang, K.; Liu, D. Smart guiding glasses for visually impaired people in indoor environment. *IEEE Trans. Consum. Electron.* **2017**, *63*, 258–266.
35. Lu, K.; Zhang, L. TBEFN: A two-branch exposure-fusion network for low-light image enhancement. *IEEE Trans. Multimed.* **2020**, *16*, 1–13. [CrossRef]
36. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 213–229.
37. Xuebin, Q.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404.
38. Mukhriddin, M.; Jeong, R.; Cho, J. Saliency cuts: Salient region extraction based on local adaptive thresholding for image information recognition of the visually impaired. *Int. Arab J. Inf. Technol.* **2020**, *17*, 713–720.
39. Bappaditya, M.; Chia, S.; Li, L.; Chandrasekhar, V.; Tan, C.; Lim, J. A wearable face recognition system on google glass for assisting social interactions. In *Asian Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 419–433.
40. Shiwei, C.; Yao, D.; Cao, H.; Shen, C. A novel approach to wearable image recognition systems to aid visually impaired people. *Appl. Sci.* **2019**, *9*, 3350.
41. Ugulino, W.C.; Fuks, H. Prototyping wearables for supporting cognitive mapping by the blind: Lessons from co-creation workshops. In Proceedings of the 2015 workshop on Wearable Systems and Applications, Florence, Italy, 18 May 2015.
42. Kumar, S.N.; Varun, K.; Rahman, J.M. Object Recognition Using Perspective Glass for Blind/Visually Impaired. *J. Embed. Syst. Process* **2019**, *4*, 31–37.
43. Fiannaca, A.; Apostolopoulous, I.; Folmer, E. Headlock: A wearable navigation aid that helps blind cane users traverse large open spaces. In Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility, Rochester, NY, USA, 20–22 October 2014.
44. Shi, Q.; Hu, J.; Han, T.; Osawa, H.; Rauterberg, M. An Evaluation of a Wearable Assistive Device for Augmenting Social Interactions. *IEEE Access* **2020**, *8*, 164661–164677.
45. Kyungjun, L.; Sato, D.; Asakawa, S.; Kacorri, H.; Asakawa, C. Pedestrian detection with wearable cameras for the blind: A two-way perspective. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020.
46. Kataoka, H.; Katsumi, H. A Wearable Walking Support System to provide safe direction for the Blind. In Proceedings of the 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Jeju, Korea, 23–26 June 2019.
47. Adegoke, A.O.; Oyeleke, O.D.; Mahmud, B.; Ajoje, J.O.; Thomase, S. Design and Construction of an Obstacle-Detecting Glasses for the Visually Impaired. *Int. J. Eng. Manuf.* **2019**, *9*, 57.
48. Ankita, B.; Laha, S.; Maity, D.K.; Sarkar, A.; Bhattacharyya, S. Smart Glass for Blind People. *AMSE J.* **2017**, *38*, 102–110.
49. Tai, S.-K.; Dewi, C.; Chen, R.-C.; Liu, Y.-T.; Jiang, X.; Yu, H. Deep Learning for Traffic Sign Recognition Based on Spatial Pyramid Pooling with Scale Analysis. *Appl. Sci.* **2020**, *10*, 6997. [CrossRef]
50. Dewi, C.; Chen, R.C.; Liu, Y.T.; Jiang, X.; Hartomo, K.D. Yolo V4 for Advanced Traffic Sign Recognition with Synthetic Training Data Generated by Various GAN. *IEEE Access* **2021**, *9*, 97228–97242. [CrossRef]

51. Chen, R.C.; Saravanarajan, V.S.; Hung, H.T. Monitoring the behaviours of pet cat based on YOLO model and raspberry Pi. *Int. J. Appl. Sci. Eng.* **2021**, *18*, 1–12. [CrossRef]
52. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5 Mb model size. *arXiv* **2016**, arXiv:1602.07360.
53. François, C. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
54. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
55. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake, UT, USA, 18–22 June 2018.
56. Xiangyu, Z.; Xinyu, Z.; Mengxiao, L.; Jian, S. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
57. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. ShuffleNet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
58. Apte, M.; Mangat, S.; Sekhar, P. YOLO Net on iOS. Available online: http://cs231n.stanford.edu/reports/2017/pdfs/135.pdf (accessed on 28 October 2021).
59. Guimei, C.; Xie, X.; Yang, W.; Liao, Q.; Shi, G.; Wu, J. Feature-fused SSD: Fast detection for small objects. In Proceedings of the Ninth International Conference on Graphic and Image Processing (ICGIP 2017), Qingdao, China, 14–16 October 2017.
60. Alexander, W.; Shafiee, M.J.; Li, F.; Chwyl, B. Tiny SSD: A tiny single-shot detection deep convolutional neural network for real-time embedded object detection. In Proceedings of the 2018 15th Conference on Computer and Robot Vision (CRV), Toronto, ON, Canada, 8–10 May 2018.
61. Wang, R.J.; Li, X.; Ling, C.X. Pelee: A real-time object detection system on mobile devices. *arXiv Prepr.* **2018**, arXiv:1804.06882.
62. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, China (Virtual), 14–19 June 2020.
63. Kim, S.; Ryu, Y.; Cho, J.; Ryu, E. Towards Tangible Vision for the Visually Impaired through 2D Multiarray Braille Display. *Sensors* **2019**, *19*, 5319. [CrossRef]
64. Cai, J.; Gu, S.; Zhang, L. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Trans. Image Process.* **2018**, *27*, 2049–2062. [CrossRef] [PubMed]
65. Chen, W.; Wang, W.; Yang, W.; Liu, J. Deep retinex decomposition for low-light enhancement. *arXiv* **2018**, arXiv:1808.04560.
66. Al-Rfou, R.; Choe, D.; Constant, N.; Guo, M.; Jones, L. Character-level language modeling with deeper self-attention. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
67. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Doll'ar, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014.
68. Kirillov, A.; Girshick, R.; He, K.; Doll´ar, P. Panoptic feature pyramid networks. *arXiv* **2019**, arXiv:1901.02446.
69. Peng, L.Y.; Chan, C.S. Getting to know low-light images with the exclusively dark dataset. *Comput. Vis. Image Underst.* **2019**, *178*, 30–42.
70. Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; Ruan, X. Learning to detect salient objects with image-level supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
71. Makhmudov, F.; Mukhiddinov, M.; Abdusalomov, A.; Avazov, K.; Khamdamov, U.; Cho, Y.I. Improvement of the end-to-end scene text recognition method for "text-to-speech" conversion. *Int. J. Wavelets Multiresolut. Inf. Process.* **2020**, *18*, 2050052-1. [CrossRef]
72. Smith, R. An overview of the Tesseract OCR engine. In Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brasil, 23–26 September 2007.
73. Abdusalomov, A.; Mukhiddinov, M.; Djuraev, O.; Khamdamov, U.; Whangbo, T.K. Automatic salient object extraction based on locally adaptive thresholding to generate tactile graphics. *Appl. Sci.* **2020**, *10*, 3350. [CrossRef]
74. Bai, J.; Liu, Z.; Lin, Y.; Li, Y.; Lian, S.; Liu, D. Wearable Travel Aid for Environment Perception and Navigation of Visually Impaired People. *Electronics* **2019**, *8*, 697. [CrossRef]
75. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
76. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
77. Padilla, R.; Passos, W.L.; Dias, T.L.B.; Netto, S.L.; da Silva, E.A.B. A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit. *Electronics* **2021**, *10*, 279. [CrossRef]
78. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.Y.; Girshick, R. Detectron2. 2019. Available online: https://github.com/facebookresearch/detectron2 (accessed on 28 October 2021).
79. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.

80. Shrivastava, A.; Gupta, A.; Girshick, R. Training region-based object detectors with online hard example mining. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
81. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.M.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
82. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollr, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
83. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake, UT, USA, 18–22 June 2018.
84. Liu, S.; Huang, D.; Wang, Y. Receptive field block net for accurate and fast object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake, UT, USA, 18–22 June 2018.
85. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
86. Zhao, Q.; Sheng, T.; Wang, Y.; Tang, Z.; Chen, Y.; Cai, L.; Ling, H. M2det: A single-shot object detector based on multi-level feature pyramid network. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Honolulu, HI, USA, 27 January–1 February 2019.
87. Bae, S.H. Object detection based on region decomposition and assembly. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Honolulu, HI, USA, 27 January–1 February 2019.
88. Zhou, X.; Zhuo, J.; Krahenbuhl, P. Bottom-up object detection by grouping extreme and center points. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
89. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
90. Abdusalomov, A.; Baratov, N.; Kutlimuratov, A.; Whangbo, T.K. An Improvement of the Fire Detection and Classification Method Using YOLOv3 for Surveillance Systems. *Sensors* **2021**, *21*, 6519. [CrossRef]
91. Zhang, P.; Wang, D.; Lu, H.; Wang, H.; Ruan, X. Amulet: Aggregating multi-level convolutional features for salient object detection. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017.
92. Chen, S.; Tan, X.; Wang, B.; Hu, X. Reverse attention for salient object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
93. Liu, N.; Han, J.; Yang, M. Picanet: Learning pixel-wise contextual attention for saliency detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake, UT, USA, 18–22 June 2018.
94. Feng, M.; Lu, H.; Ding, E. Attentive feedback network for boundary-aware salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
95. Zeng, Y.; Zhuge, Y.; Lu, H.; Zhang, L.; Qian, M.; Yu, Y. Multi-source weak supervision for saliency detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
96. Wang, T.; Borji, A.; Zhang, L.; Zhang, P.; Lu, H. A stagewise refinement model for detecting salient objects in images. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017.
97. Wu, Z.; Su, L.; Huang, Q. Cascaded partial decoder for fast and accurate salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
98. Liu, J.; Hou, Q.; Cheng, M.; Feng, J.; Jiang, J. A simple pooling-based design for realtime salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
99. Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; Jagersand, M. Basnet: Boundaryaware salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
100. Zhang, Z.; Zhang, C.; Shen, W.; Yao, C.; Liu, W.; Bai, X. Multi-oriented text detection with fully convolutional networks. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
101. Yao, C.; Bai, X.; Sang, N.; Zhou, X.; Zhou, S.; Cao, Z. Scene text detection via holistic, multi-channel prediction. *arXiv* **2016**, arXiv:1606.09002.
102. Shi, B.; Bai, X.; Belongie, S. Detecting oriented text in natural images by linking segments. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
103. He, W.; Zhang, X.Y.; Yin, F.; Liu, C.L. Deep direct regression for multi-oriented scene text detection. In Proceedings of the IEEE Conference Computer Vision, Venice, Italy, 22–29 October 2017.
104. Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; Liang, J. EAST: An efficient and accurate scene text detector. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
105. Long, S.; Ruan, J.; Zhang, W.; He, X.; Wu, W.; Yao, C. TextSnake: A flexible representation for detecting text of arbitrary shapes. In Proceedings of the European Conference Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
106. Deng, D.; Liu, H.; Li, X.; Cai, D. Pixellink: Detecting scene text via instance segmentation. In Proceedings of the Thirty-Second AAAI Conference Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
107. Wang, F.; Zhao, L.; Li, X.; Wang, X.; Tao, D. Geometry-aware scene text detection with instance transformation network. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
108. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep structured output learning for unconstrained text recognition. *arXiv* **2014**, arXiv:1412.5903.

109. Shi, B.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. Robust scene text recognition with automatic rectification. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
110. Shi, B.; Bai, X.; Yao, C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 2298–2304. [CrossRef] [PubMed]
111. Lee, C.Y.; Osindero, S. Recursive recurrent nets with attention modeling for OCR in the wild. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
112. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Reading text in the wild with convolutional neural networks. *Int. J. Comput. Vis.* **2016**, *116*, 1–20. [CrossRef]
113. Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. Aster: An attentional scene text recognizer with flexible rectification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2035–2048. [CrossRef]
114. Cheng, Z.; Bai, F.; Xu, Y.; Zheng, G.; Pu, S.; Zhou, S. Focusing attention: Towards accurate text recognition in natural images. In Proceedings of the IEEE Conference Computer Vision, Venice, Italy, 22–29 October 2017.
115. Yang, K.; Bergasa, L.M.; Romera, E.; Cheng, R.; Chen, T.; Wang, K. Unifying terrain awareness through real-time semantic segmentation. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018.
116. Mancini, A.; Frontoni, E.; Zingaretti, P. Mechatronic system to help visually impaired users during walking and running. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 649–660. [CrossRef]
117. Patil, K.; Jawadwala, Q.; Shu, F.C. Design and construction of electronic aid for visually impaired people. *IEEE Trans. Hum.-Mach. Syst.* **2018**, *48*, 172–182. [CrossRef]
118. Al-Madani, B.; Orujov, F.; Maskeliūnas, R.; Damaševičius, R.; Venčkauskas, A. Fuzzy logic type-2 based wireless indoor localization system for navigation of visually impaired people in buildings. *Sensors* **2019**, *19*, 2114. [CrossRef]

entropy

MDPI

Article

Efficient Multi-Object Detection and Smart Navigation Using Artificial Intelligence for Visually Impaired People

Rakesh Chandra Joshi [1], Saumya Yadav [1], Malay Kishore Dutta [1,*] and Carlos M. Travieso-Gonzalez [2]

[1] Centre for Advanced Studies, Dr. A.P.J. Abdul Kalam Technical University, Lucknow 226031, India; rakeshchandraindia@gmail.com (R.C.J.); saumyay.15@gmail.com (S.Y.)

[2] Institute for Technological Development and Innovation in Communications (IDeTIC), University of Las Palmas de Gran Canaria (ULPGC), 35017 Las Palmas de G.C., Spain; carlos.travieso@ulpgc.es

* Correspondence: malaykishoredutta@gmail.com

Received: 8 July 2020; Accepted: 22 August 2020; Published: 27 August 2020

Abstract: Visually impaired people face numerous difficulties in their daily life, and technological interventions may assist them to meet these challenges. This paper proposes an artificial intelligence-based fully automatic assistive technology to recognize different objects, and auditory inputs are provided to the user in real time, which gives better understanding to the visually impaired person about their surroundings. A deep-learning model is trained with multiple images of objects that are highly relevant to the visually impaired person. Training images are augmented and manually annotated to bring more robustness to the trained model. In addition to computer vision-based techniques for object recognition, a distance-measuring sensor is integrated to make the device more comprehensive by recognizing obstacles while navigating from one place to another. The auditory information that is conveyed to the user after scene segmentation and obstacle identification is optimized to obtain more information in less time for faster processing of video frames. The average accuracy of this proposed method is 95.19% and 99.69% for object detection and recognition, respectively. The time complexity is low, allowing a user to perceive the surrounding scene in real time.

Keywords: artificial intelligence; assistive systems; computer vision; deep learning; machine learning; object recognition; visually impaired person; YOLO-v3

1. Introduction

Vision impairment is one of the major health problems in the world. Vision impairment or vision loss reduces seeing or perceiving ability, which cannot be cured through wearing glasses. Navigation becomes more difficult around places other than the visually impaired person's own home or places that are not familiar. Vision impairment is classified into near and distance vision impairment. In near vision impairment, vision is poorer than M.08 or N6, even after correction. Distance vision impairment is classified into mild, moderate, severe, and blindness based on visual acuteness, when it is worse than 6/12, 6/18, 6/60, and 3/60, respectively [1]. About 80% of people who suffer from visual impairment or blindness belong to middle- and low-income countries, where they cannot afford costly assistive devices. The problem arises due to an increase in age or population [2]. Vision impairment can be due to many reasons such as uncorrected refractive errors, age-related eye problems, glaucoma, cataracts, diabetic retinopathy, trachoma, corneal opacity, or unaddressed presbyopia [3].

Apart from medical treatment, people use various aids for rehabilitation, education, social inclusion, or work. A white cane is used by visually impaired people around the world. The length of the cane is

directly proportional to the range of touch sensation or the detection of obstacles. Guide dogs are also used as walking assistance, where the dog makes the user aware of obstacles or for stepping up and down. However, guide dogs are unable to give directions in complex cases. People also make use of GPS-equipped assistive devices, which help with navigation and orientation for a particular position. These kinds of devices are accurate in terms of location, but are ineffective in case of obstacle avoidance and object identification. Echolocation [4] is another technique used by blind people in which echoes of sounds made by simple mouth clicks are used to detect silent objects in front of them.

Braille helps a visually impaired or blind person to obtain information, but it is limited to people who have knowledge of it. Information in braille characters can be installed in most places, but it is not practical to install it everywhere and convey full information. Currency notes also feature the tactile marks with raised dots to allow the person to identify the banknote. However, these tactile marks vanish after some time, and then it is not easy for a blind person to differentiate between banknotes. Refreshable braille displays, screen magnifiers, and screen readers are also used to obtain information while using computer or mobile systems.

Visually impaired people use Electronic Travel Aids (ETA) to detect obstacles and identify services to provide safe and informative navigation. A hardware-based robotic cane is proposed for assistance in walking. It has an omnidirectional wheel assisted with a high-speed processing controller using a LAM-based linearization system with a non-linear disturbance observer. It maintains the balance of the person and reduces the risk of falling. The length, cost, and weight are other parameters, which can be optimized for better support [5]. An Electronic Mobility Cane (EMC) is designed for vision rehabilitation of visually impaired people to provide assistance and detect obstacles [6], where a logical map is constructed to obtain information about the surrounding environment. Output information is conveyed in the form of audio, vibration, or voice. A haptic device such as a short cane with smart sensors is proposed in [7] to provide information about obstacles. Different ultrasonic sensors are associated with providing the same stimuli of a traditional stick without touching obstacles. Another multi-sensor ETA device proposed in [8] has a pair of eyeglasses that guide people in a safe and efficient manner using ultrasonic and depth sensors to provide a navigational guide to the visually impaired person. A cane robot is proposed in [9] for assistance and fall prevention. Some object-recognition techniques are based on Quick Response (QR) codes and barcodes to identify different types of objects at various places such as in shopping malls, etc. [10,11], which requires an advanced infrastructure.

A smart cane and obstacle-detection system for visually impaired people with multiple sensors has been designed with model-based state-feedback control [12]. A linear–quadratic regulator (LQR) based controller is also integrated for the optimization of an actuator's control actions along with position tracking. A white-cane system that is composed of IC tags is designed in [13], which supports independent walking of visually impaired people in indoor space. Colored lines on the floor for navigation is sensed by the cane, and information to reach the destination is given with vibrations and voice prompts. An intelligent system is proposed in [14] which contains map information for independent navigation walking for visually impaired people while walking in indoor space. The color on the floor is recognized through the one-chip microprocessor and Radio Frequency Identification (RFID) system. Tactile zooming is discussed in [15] for graphics, which make it easy for visually impaired people to obtain information by magnification. Navigation assistance for the visually impaired (NAVI) is developed to assist through sound commands using an RGB-D camera [16]. Automatic recognition of clothing patterns is done in [17] for visually impaired people, which can identify 11 colors of cloth. The system consists of a mounted camera on goggles, a microphone, a Bluetooth ear piece, and a computer. A wearable device is designed using a haptic strap, sensor belt, and vibration motors to detect different types of obstacles and allow for safe navigation [18]. An object-detection method is proposed in [19], which is based on a deformable grid (DG), which depends upon the motion of an object and can detect the risk of collision.

In [20], an automatic quantization algorithm is developed for deep convolutional neural network (DCNN)-based object detection that uses a smaller number of bits and reduces the hardware cost compared to traditional methods. The main challenge in developing an assistive framework using a CNN architecture is to increase the accuracy for the classification task while maintaining an acceptable computational workload [21]. A DEEP-SEE FACE framework-based assistive device is introduced for improving the cognition and communication of visually impaired people by recognizing known faces and differentiating them from unknown faces [22]. A light detection and ranging (LIDAR)-assisted system is proposed in [23] to obtain spatial information through the stereo sound of different pitches. A tracking system for indoor and outdoor navigation using computer vision algorithms and dead reckoning to help visually impaired people has been implemented in a smartphone [24]. An electronic device for the automatic navigational assistance of a visually impaired person, named NavCane, is developed in [25] for the obstacle detection of various types with different types of sensors at different positions of a white cane. SUGAMAN [26] is a framework developed to describe floor plans using proximity-based grammar and learning annotation. It utilizes the text information in floor plan images to develop proper navigation and obstacle avoidance for visually impaired persons. A wearable deep learning-based drug pill recognition system for the improvement in safety to visually impaired people using medicines works by reducing the risk of taking incorrect drugs [27]. An assistive device to help visually impaired people using partial visual information and machine learning techniques which enables semantic categorization for the classification of obstacles in front of them achieves a highest accuracy of 90.2% [28].

Science and engineering can make technical interventions to the lives of visually impaired people in making them independent to navigate and perceive objects around them. Many devices have been proposed to assist a visually challenged person, but most of the devices focus either on object detection through computer vision or on obstacle detection through different sensors, such as GPS, distance sensors, etc. However, the effective utilization of sensor-based technologies and computer vision could result in a highly efficient and supportive device, to make them aware of the surroundings.

The main contribution of the proposed work is to design an artificial intelligent fully automated assistive technique for visually impaired people to perceive the objects in the surrounding and provide obstacle-aware navigation, where auditory inputs are given to users in real-time. Images of objects that are highly relevant in the lives of the visually challenged are trained using deep learning neural networks. Augmentation and manual annotation are performed on the dataset to make the system robust and free from overfitting. Both sensors and computer-vision based techniques are integrated to provide convenient a travel-aid to visually impaired people, through which a person can perceive multiple objects, detect obstacles and avoid collisions.

The whole framework is standalone and designed for low-cost processors, so that the visually impaired can use it properly without internet connectivity. The detected output is also optimized to ensure faster frame processing and a greater extraction of information—i.e., count of objects—in a shorter time period. The proposed system can easily differentiate between obstacles and known objects. The proposed methodology can make a significant contribution to assist visually impaired people compared to previously developed methods, which were only focused on obstacle detection and location tracking with the help of basic sensors without use of deep learning. With the use of the proposed methodology, users can interpret more about the things going around and receive obstacle-aware navigation as well.

The rest of the paper is organized as follows: First, the methodology is explored in Section 2, followed by the experimental results presented in Section 3. Finally, our conclusions and ideas for future work are discussed in Section 4.

2. Methodology

In this section, the whole process is explained to provide navigation assistance to visually impaired people, which consists of the preparation and pre-processing of the dataset, augmentation,

annotation and dataset training on the deep-learning model. The block diagram for proposed methodology is represented in Figure 1.

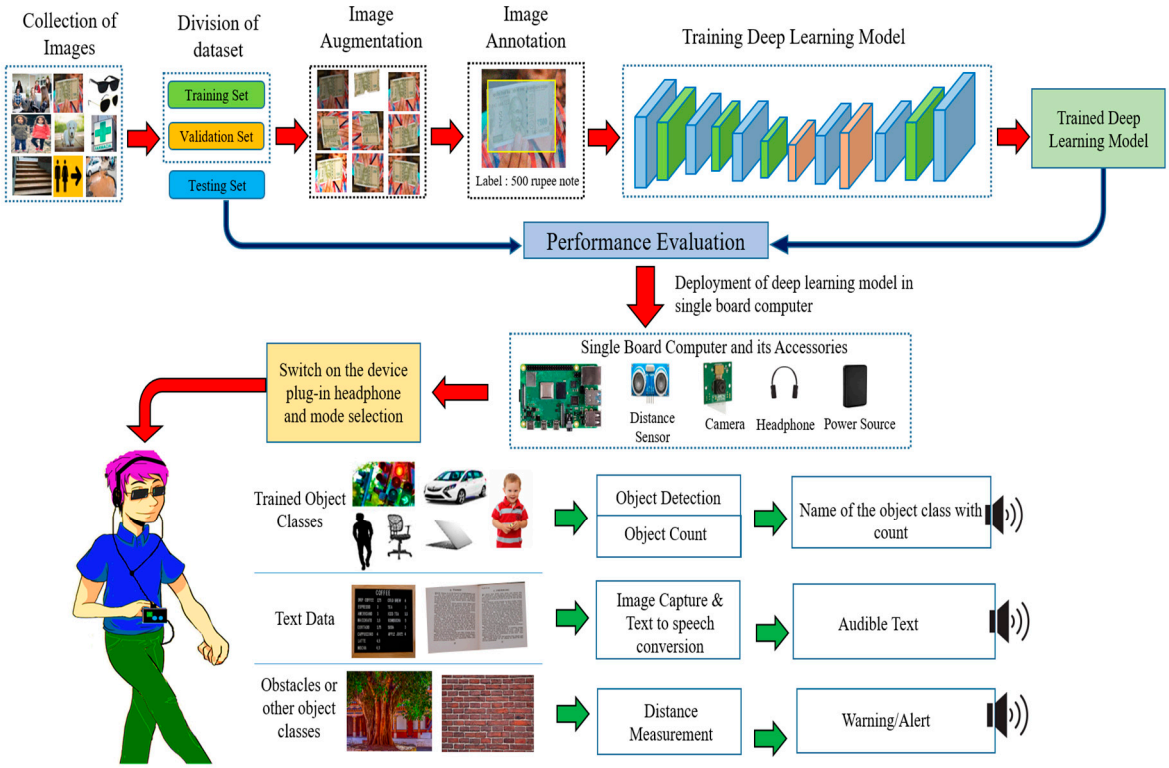


Figure 1. Block Diagram for proposed methodology.

2.1. Dataset for Visual Impaired People

Many datasets are available for object detection, such as PASCAL [29], CIFAR 10 [30], IMAGENET [31], SUN [32] and MS COCO [33] but these contain limited classes from the perspective of assisting visually impaired persons. Thus, there is a need to add more objects in existing datasets so that they can help visually disabled persons to be socially independent. A survey was conducted in visually disabled schools and colleges to select more relevant objects to train a deep-learning model. The dataset was generated from multiple sources and devices, in different sizes and pixels. Various lighting conditions and capturing angles were used to make more variations in the collected dataset. The banknote/currency notes were also included in the dataset, to perform cash transactions with ease. Thereafter, those images which had less than 10% area of the targeted object or any deformities—such as flickering, blur or noise to more than an acceptable extent—were eliminated. After that, augmentation variants were applied to the captured and collected images.

2.2. Image Augmentation

All collected images were then augmented to resist the trained model from overfitting and to perform more robust and accurate object detection for visually impaired persons. Various augmentation techniques, such as rotation at different angles, skewing, mirroring, flipping, brightness levels, noise levels, and a combination of these techniques, was used to enrich the dataset to many folds, shown in Figure 2. As banknotes are also a part of daily life, various images of different denominations of banknotes were collected and augmented before training the neural network to recognize banknotes efficiently and accurately.

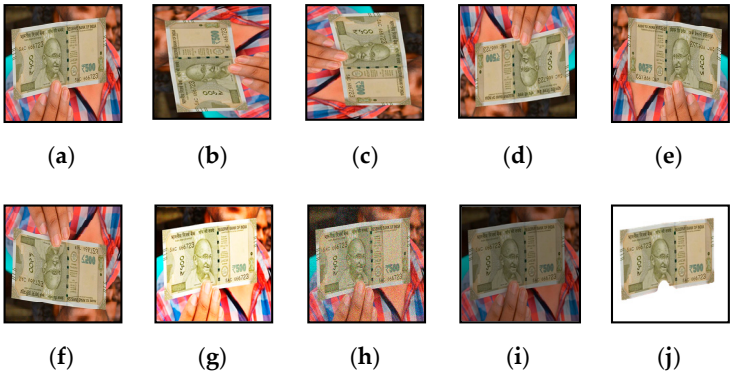


Figure 2. Different image augmentation techniques on acquired images: (**a**) Original Image; (**b**) 90° left rotation; (**c**) 90° right rotation; (**d**) 180° Horizontal rotation; (**e**) Horizontal Flip; (**f**) Vertical Flip; (**g**) Increased Brightness; (**h**) Addition of noise; (**i**) Low contrast; (**j**) Background Removal.

2.3. Image Annotation

All images were annotated manually with the LabelImg tool and the bounding box was made around the object without taking extra unnecessary areas. The information about the images, such as the size of the image, size, and position of the bounding box or bounding boxes (in case of multiple instances or multiple objects in the same image), were recorded and saved into the ".xml" format. Once the images were annotated, the respective annotation files were also generated. The final dataset, that consists of annotated images and respective annotation files, was divided into two sets—training and validation. Then, the YOLO-v3 model is trained with the generated dataset either through transfer learning or with direct training.

The transfer learning method requires a pre-trained model and it will be beneficial when a similar dataset is already trained over this model and respective generated trained model files will be used for transfer learning. Due to this, weight adjustment takes less time compared to the case when training the dataset for the first time. As weight adjustment and loss in each convolving layer reduce in a shorter time, the transfer learning method can also be used to retrain the dataset when the training got abrupt due to any reasons.

2.4. Dataset Training on Deep-Learning Model

In the YOLO-based object detection [34], the given image was divided into grids of S × S where S = a number of grid cells in each of the axes. There, each unit of the grid was accountable to detect the targets which were getting into it. Then, a corresponding confidence score was predicted for the B number of bounding boxes by each of the grid units. The confidence score represents the similarity with the desired object and maximum likelihood represents a higher confidence score of the corresponding object. In other words, it defines the presence and absence of any object class in the image. In the same way, if the object did not contain the desired object, the confidence score would be zero. If the object was contained by the predicted bounding box, then the confidence score would be calculated by the interaction in between both bounding boxes, i.e., predicted and ground truth represented by the Interaction over Union (IOU). Equation (1) is used to calculate the confidence score in the given input image.

$$CS = P_r(Obj) * IOU_{Groundtruth}^{Predicted} \quad (1)$$

where, CS = Confidence Score, $P_r(Obj)$ represents the probability of the object and $IOU_{Groundtruth}^{Predicted}$ represents the IOU of predicted and ground truth bounding boxes.

Loss function for YOLO architecture is given by Equation (2).

$$\begin{aligned} \text{Loss} = \lambda_{coord} & \sum_{i=0}^{S^2}\sum_{j=0}^{B} \mathbb{1}_{ij}^{obj}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2] \\ & +\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} \mathbb{1}_{ij}^{obj}[\left(\sqrt{w_i}-\sqrt{\hat{w}_i}\right)^2+\left(\sqrt{h_i}-\sqrt{\hat{h}_i}\right)^2] \\ & +\sum_{i=0}^{S^2}\sum_{j=0}^{B} \mathbb{1}_{ij}^{obj}(C_i-\hat{C}_i)^2 \\ & +\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B} \mathbb{1}_{ij}^{noobj}(C_i-\hat{C}_i)^2+\sum_{i=0}^{S^2}\mathbb{1}_{i}^{obj}\sum_{c\in classes}(p_i(C)-\hat{p}_i(c))^2 \end{aligned} \tag{2}$$

where, $\mathbb{1}_{ij}^{obj}$ denotes the j^{th} bounding box predictor in the i^{th} cell, which is also responsible for prediction, and $\mathbb{1}_{i}^{obj}$ denotes if the object appears in cell i.

YOLO-v3 [35] is an upgraded version of YOLO and YOLOv2 [36] for object detection in real-time and accurately. YOLO-v3 uses logistic regression is utilized instead of Softmax to predict the objectness score for each bounding box. Thus, multi-label classification and class prediction can be performed using YOLO-v3. Feature Pyramid Networks (FPN) in YOLO-v3 makes three predictions for each location of the input frame and features are extracted from each prediction, which include boundary box and objectness scores.

Darknet-53 is used as a feature extractor in YOLO-v3 that is composed of 53 convolutional layers. It runs at the highest measured floating-point operation speed, which indicates that the network is more successful when applying GPU resources [37]. The network architecture of YOLO-v3 with Darknet-53 is shown in Figure 3.

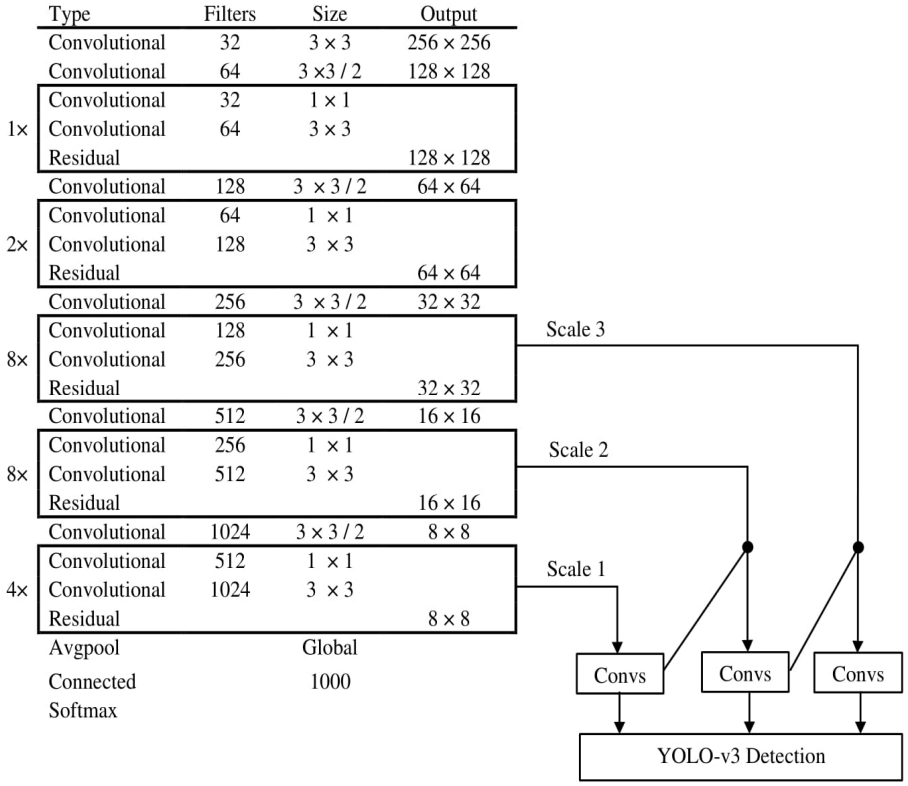


Figure 3. Network Structure of YOLO-v3.

In the training neural network, the predictions were made through the following Equations (3)–(6):

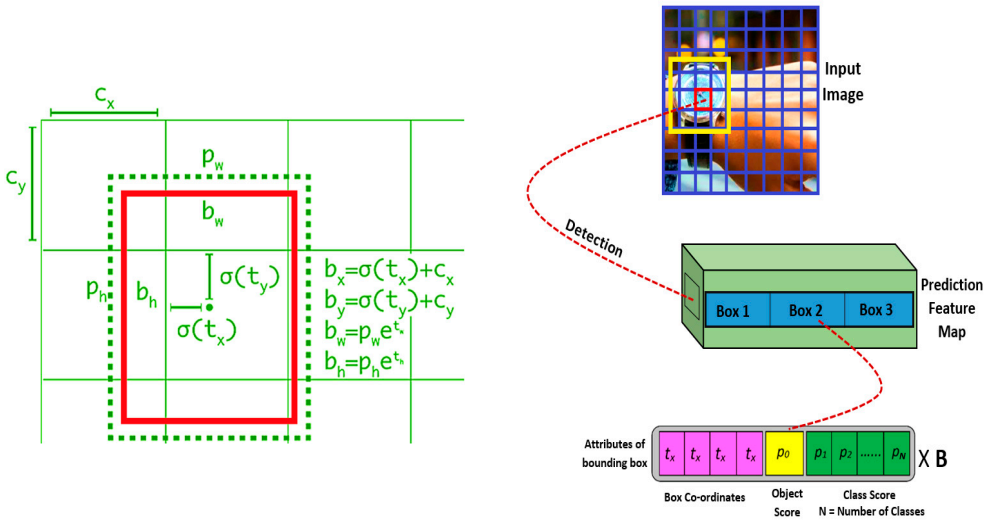
$$b_x = \sigma(t_x) + c_x \ldots \tag{3}$$

$$b_y = \sigma(t_y) + c_y \tag{4}$$

$$b_w = p_w \, e^{tw} \tag{5}$$

$$b_h = p_h \, e^{th} \tag{6}$$

where (t_x, t_y, t_w, t_h) are four coordinates that were predicted for each of the bounding boxes. (c_x, c_y) is the cell offset from the top left image corner, and (p_w, p_h) are the width and height of the bounding box prior. The diagrams for bounding box prediction and object detection with the training model are shown in Figure 4.



(**a**) (**b**)

Figure 4. (**a**) Bounding Box Prediction; (**b**) Object detection with YOLO-v3.

Once the CNN was trained with the dataset, the final trained model was equipped in the object detection framework. A live video feed was associated to the framework and image frames were subsequently captured. Captured frames were pre-processed and fed into the trained model, and if any object which was trained with the model was detected, a bounding box was drawn around that object and a respective label was generated for that object. Once all objects were detected, the text label was converted into speech, or a respective audio label recording was played, and subsequently, the next frame was processed. The Algorithm 1 elaborates the steps of object detection for a visually impaired person after the training of the dataset is as follows:

Algorithm 1. Object detection for visually impaired person after training of dataset.

Input: Captured image from the Camera
Output: Audio for the label of the detected object
Step 1: *Save the captured image, I*
Step 2: *Pre-processing of the image*
*Resize the image in dimensions w*h*
where, w = number of pixels in the x-axis
h = number of pixels in the y-axis
Increase the contrast value for I
Step 3: *Load the trained deep-learning model and its parameters*
Step 4: *Image I is processed with the deep-learning model*
detections = detectObjectsFromImage(input_image = I)
Step 5: *Save processed output image, O*
for *detections*
Bounding Box prediction (b_x, b_y, b_w, b_h)
Percentage probability of the object
Label, l = name of the detected object
Text to speech conversion for l
end

The proposed module consists of a DSP processor with a distance sensor, camera, and power supply. Speakers or headphones are associated with the DSP processor to perceive predictions as an audio prompt.

Output information optimization was further performed to increase the robustness of the system. If an object is detected in the captured image frame, equivalent audio is played after the detection of an object to convey information to the user. Thus, the information transmission time will increase with an increase in the number of the objects in current image frame and cause a delay to processing the next frame. This problem is not discussed in many research articles where such work is conducted. Frame processing time in blind assistive devices is different for a normal human and visually impaired persons. In the case of assistance for visually impaired people, the frame processing time also includes the time necessary to convey detection information as audio or vibrations. Thus, even though the machine learning model processes the frames in real-time, it takes a lot of time to process the next frame, as it has a dependency on the number of the objects present in the current frame and the length of the name of object. For example, the time taken to pronounce "car" is less than that required to pronounce "fire extinguisher". Thus, three steps have been taken to deal with these kinds of problems. First, all audio files for the name of objects label are optimized such that there is no silence in recording, except the space in between two words. Recording playback speed is increased to the extent that it still sounds clear and understandable.

Second is the case where the same kind of objects exist multiple times in the captured frame. For example, in a case where 5 people are present in the scene, the conventional system will take the equivalent of five times to prompt the word "person", or more (because of a time gap in between pronouncing two words). To optimize this, the object counter is added with a trained model that counts the number of objects of same category in current image frame, processes it, and conveys a piece of audio information with both "number of objects" and "name/label of an object". Thus, the time taken to prompt "person" five times is reduced to "5 person". Consequently, the time taken to process the next frame is reduced, which results in a smaller instant of time to convey information.

Third is the case in which a number of multiple objects of various categories are present in the captured scene, which can require a considerably longer time to convey audio information to the user. To deal with this issue, the number of object categories is limited to three, but can be extended to five object classes for indoor circumstances. This means that even though the number of objects which were detected is higher, the system will convey the information of all objects of only the first

three categories and then process the next frame. With these three improvements in information transmission, the processing time between two frames is reduced. A flow chart diagram for the optimized information transmission with the object counter is shown in Figure 5.

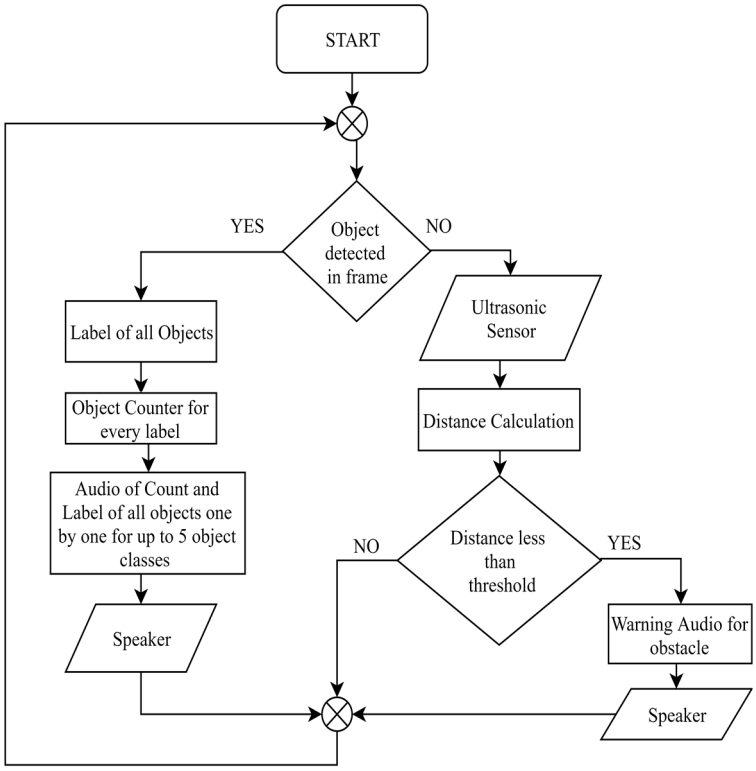


Figure 5. Information Optimization and Object–Obstacle Differentiation.

If none of the trained objects are detected in the captured frame, then it will calculate the distance through the ultrasonic sensors. If the calculated distance is less than the threshold, then it will be considered as an obstacle. Otherwise, if the calculated distance is more than the threshold, then the next image frame will be captured and processed.

All previous inventions and research works for blind or visually impaired people which use ultrasonic sensors to detect obstacles define their range and play a warning sound whenever an obstacle comes across the sensor. When the calculated distance through the ultrasonic sensor is below the threshold value, the device makes an acoustic warning or vibrates, but it can be irritating for a visually impaired person who is standing in a crowd and repeatedly listening same prompt or continuous vibrations. So, one of the objectives of the proposed system is to differentiate between trained objects and obstacles.

The system first analyses the current frame for object detection—if an object is detected which means the object is in front of the device, then there is no need for searching another obstacle. If no objects are identified in present frame, then it takes input from an ultrasonic sensor regarding the distance from the object, and if the calculated distance is less than the threshold, then it treats object as an obstacle and warns the person through an auditory message, as shown in flow chart in Figure 5. A vibration motor can also be associated so that it can vibrate at that instance. However, an auditory response is good in many respects, as it does not annoy a person unlike vibrations, and the power requirement is less compared to a vibration motor.

Different modes are designed in the device to provide wider assistance such as indoor, outdoor or text-reader mode. The activity diagram for the working of the assistive framework is illustrated in Figure 6.

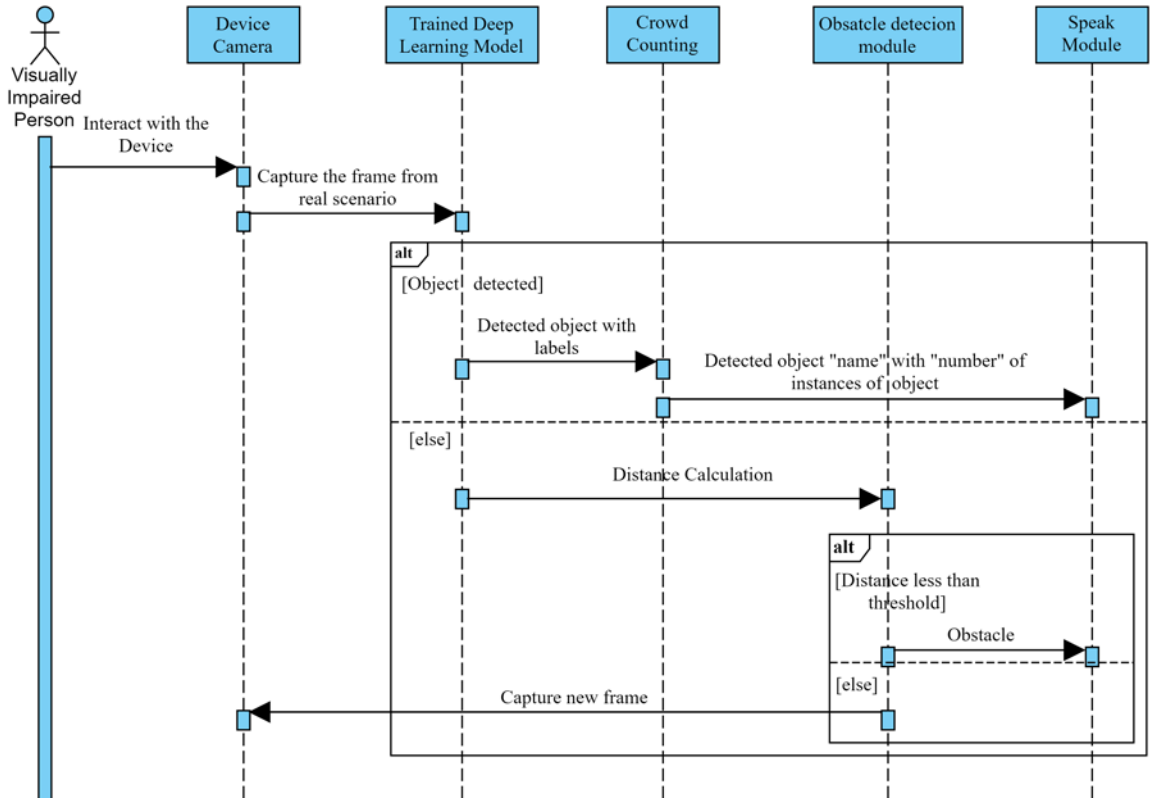


Figure 6. Activity Diagram for the proposed assistive approach for visually impaired people.

The indoor mode has a smaller threshold value for the distance of obstacle compared to the outdoor mode. Outdoor mode also has an image enlarge function, so that far objects can be detected early. For example, a car at far distance can be easily detected when enlarged, because after enlarging the image, the number of pixels is increased, and it becomes an easy task to detect that car. This feature provides an audio prompt when the object is at distance, and helps user to be alert to their surroundings, as early detection is crucial in case the user is outside and especially in those scenarios such as when a car is coming towards the person. The text reader mode can be used efficiently where the user has a necessity to read, such as when reading a book, a restaurant menu, etc. To read text, Optical Character Recognizer (OCR) is used after preprocessing the input image frame. Face recognizer can also be associated with the device, where users can identify known persons and family members, which will help in them to be social and secure.

3. Experiments and Results

The hardware specifications of training device are i-9 processor, NVIDIA Tesla K80 GPU, having 2496 CUDA cores, 12 GB GDDR5 VRAM. The system is made for hundred objects of different classes. The model is also trained to perform banknote detection and recognition to help in daily business transaction-related activities along with other object detection and navigation assistance for visually impaired people. The whole set-up is implemented in the single board DSP processor and has specifications of 64-bit, quad-core, and 1.5 GHz, as well as 4 GB SDRAM. The 8-megapixel camera used can capture images of 3280 × 2464 pixels with a fixed focus lens.

In total, 650 images of each class were collected and, out of those, 150 images were kept separated for the testing set. The remaining 500 images from each training class were divided into a ratio of 7:3 for training and validation set, respectively. After completing augmentation, the dataset in the training and validation set was increased by 10 times the initial set of images, which resulted in a wide variety of images. The number of images in the given dataset is given in Table 1. Augmentation induces the

robustness in the training model. The Deep Learning model is trained with the dataset at an initial learning rate of 10^{-3}. Training is performed until the loss is reduced and becomes saturated at a certain epoch. In between the training processes, the trained model files for lower loss can be used to test the detection and recognition performance of the system to conduct a subsequent analysis of the trained system. If a trained model performs poorly with lower loss model files, either the dataset should be increased, or various augmentations should be performed on an existing dataset.

Table 1. Number of images in each class of the collected dataset.

Total Images in Each Object Class	Original Dataset		After Augmentation		Test Set
	Training Set	Validation Set	Training Set	Validation Set	
650	350	150	3500	1500	150

The model file after training on different object classes is tested on a real-time live video feed along with images left for the testing dataset. Table 2 is prepared for the analysis of object detection and recognition accuracy of proposed system. An average accuracy of 95.19% is achieved for object detection and the average recognition accuracy is 99.69%. The results signify that once the object is detected it will be classified properly among the list of object classes, which were trained on a prepared dataset. As objects are trained regressively, the high threshold will also withstand with the accuracy.

Table 2. Performance analysis of proposed model on most relevant objects.

Objects	Total Testing Images	Correctly Detected	Detection Accuracy (%)	Correctly Recognized	Recognition Accuracy (%)
Person	150	148	98.67	148	100.00
Car	150	146	97.33	145	99.32
Bus	150	144	96.00	144	100.00
Truck	150	143	95.33	141	98.60
Chair	150	147	98.00	146	99.32
TV	150	140	93.33	140	100.00
Bottle	150	148	98.67	148	100.00
Dog	150	145	96.67	144	99.31
Fire hydrant	150	146	97.33	146	100.00
Stop Sign	150	149	99.33	147	98.66
Socket	150	143	95.33	143	100.00
Pothole	150	129	86.00	128	99.22
Pharmacy	150	141	94.00	139	98.58
Stairs	150	139	92.67	139	100.00
Washroom	150	145	96.67	145	100.00
Wrist Watch	150	140	93.33	139	99.29
Eye glasses	150	141	94.00	141	100.00
Cylinder	150	131	87.33	131	100.00
10 ₹ Note	150	141	94.00	141	100.00
20 ₹ Note	150	148	98.67	148	100.00
50 ₹ Note	150	143	95.33	143	100.00
100 ₹ Note	150	140	93.33	140	100.00
200 ₹ Note	150	144	96.00	144	100.00
500 ₹ Note	150	140	93.33	140	100.00
2000 ₹ Note	150	149	99.33	149	100.00
Average			**95.19%**		**99.69%**

Confusion matrix is another parameter that can be utilized to check the performance of object detection and recognition on a set of test data whose true values are known. It checks whether the system is capable of differentiating between the two classes of objects after the detection. The higher values in the respective classes show the high differentiation between the two classes. As the similarity in banknotes is greater, confusion matrix for currency notes is shown in Figure 7, taking the highest percentage prediction into consideration. Differentiation between two classes is tougher when two classes are almost similar in appearance. For example, if a banknote of INR 2000 is tested in a folded position and digits are focused, then there could be confusion between INR 20, 200 or 2000. In such cases, the model trained using the dataset predicts the banknote denomination for the captured picture, but it will give a higher value of detection percentage to true value of banknote as it is also trained with the texture of notes. Thus, the overall resemblance with the true value of banknotes will be higher, which can be concluded from the confusion matrix.

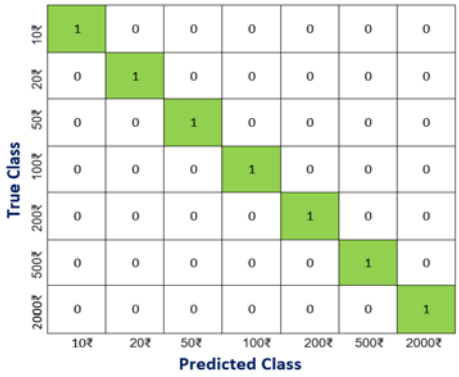


Figure 7. Confusion Matrix.

The confusion matrix is prepared for a threshold of 0.5; because of this, if the captured image is not proper, there may be chances that image shows some similarity with other banknotes along with the actual currency note. This issue can be easily eliminated by increasing the threshold value or by considering only the highest label prediction probability. Thus, if there is a currency detection mode in the device, that mode must have a higher object detection threshold value than other modes to avoid such ambiguity.

Once the performance testing is complete, the trained model is loaded onto a small DSP processor and equipped with ultrasonic sensors to detect the obstacles. Results for different object classes in different scenarios are shown below in Figure 8. Trained deep-learning models can detect and recognize the object correctly, which proves the accuracy and robustness of the proposed system.

Figure 8. Few results after object detection and recognition.

Different approaches for object classification and object detection were also tested in given datasets, such as VGG-16, VGG-19 and Alexnet. The testing accuracy and processing time for a single image frame are given in Table 3.

Table 3. Testing accuracy and frame processing time for proposed and other methods.

Methods	Testing Accuracy	Frame Processing Time
AlexNet [38]	83.39	0.275 s
VGG-16 [39]	86.80	0.53 s
VGG-19 [40]	90.21	0.39 s
YOLO-v3	95.19	0.1 s

Information optimization is performed to get more information in a shorter time duration. The time-domain analysis of the proposed system is given in Tables 4 and 5. Table 4 explains the parameters and average time taken to perform each step, whereas Table 5 explains object detection in different scenarios, such as single object single instances, single object multiple instances, multiple object single instances, and multiple object multiple instances. All the time parameters are given for a single board DSP processor without GPU support.

Table 4. Average time taken for different parameters.

Parameters	Average Time Taken (s)
Object Detection in single frame with GPU	0.1
Object Detection in single frame in single board DSP processor without GPU	0.3
Average time of Audio for name of object	0.4
Average time of Audio for count of object	0.2
Average time of Audio for name of object with count	0.6

Table 5. Processing time of each frame in different condition in single board computer.

Number of Object Class	Number of Instances of Each Object	Total Number of Objects in Frame	Average Time Taken for Object Detection (s)	Average Time of Audio Prompt (s)	Total Time to Process Single Frame (s)
0	0	0	0.3	0	0.3
1	1	1	0.3	0.4	0.7
1	2	2	0.3	0.6	0.9
1	5	5	0.3	0.6	0.9
2	1	2	0.3	0.4 + 0.4	1.1
2	2	4	0.3	0.6 + 0.6	1.5
3	5	15	0.3	0.6 + 0.6 + 0.6	2.1
4	1	4	0.3	0.4 + 0.4 + 0.4 + 0.4	1.9
4	5	20	0.3	0.6 + 0.6 + 0.6 + 0.6	2.7
5	1	5	0.3	0.4 + 0.4 + 0.4 + 0.4 + 0.4	2.3
5	3	15	0.3	0.6 + 0.6 + 0.6	2.1
5	5	25	0.3	0.6 + 0.6 + 0.6 + 0.6 + 0.6	3.3
5	10	50	0.3	0.6 + 0.6 + 0.6 + 0.6 + 0.6	3.3

Resources are used in an optimized way to reduce energy consumption. Ultrasonic sensors derive power only when the objects are not present in a captured scene. As the model is trained with most of those objects that it comes across in daily life, there is a smaller probability that the ultrasonic sensor will be used, apart from a case where the user is within a closed space with a distance less than the threshold.

The device is programmed to work in a fully automatic manner to perform object recognition and obstacle detection. For switching in between different modes, a person must swipe their hand in front of the device, which can be sensed by ultrasonic sensors to perform mode-switching. Device instructions can also be made multi-lingual by just recording the instructions in other languages. As it does not depend on a computer language interpreter, instructions can also be made for local dialect or language for which proper recordings are not yet available. The device can work in real-time scenarios, as the processing time for object detection is a few milliseconds. The higher the processor, the greater the number of frames per seconds that can be processed.

If a user wants to record image frames, which came across the device, it can be stored in subsequent frames. These frames can also help to construct a proper dataset and to approach the challenging scenario, which can be dealt with to develop much more robust devices. Above all, the whole system is standalone and needs no internet connection to perform object detection and safe navigation.

After training the collected dataset with various image augmentation techniques and multi-scale detection functionality of trained deep neural network, the proposed framework is able to detect objects in different scenarios, such as low illumination, different viewing angles, and various scale objects. The proposed system can work universally in the existing infrastructure which has been used before by visually impaired people.

Proposed work is also compared with the other works in the domain of assistance for the visually impaired and is shown in Table 6.

Table 6. Comparison with state-of-the-art methods.

Method	Components	Dataset	Result	Coverage Area	Connection	Cost
Hoang et al. [41]	Mobile Kinect, laptop Electrode matrix, headphone and RF transmitter	Local dataset	Detect obstacle and generate audio warning	Indoor	Offline	High
Bai et al. [8]	Depth camera, glasses, CPU, headphone and ultrasonic sensor	Not included	Obstacle Recognition and audio output	Indoor	Offline	High
Yang et al. [42]	Depth Camera on Smart glass, Laptop, and headphone	ADE20, PASCAL, and COCO	Obstacle Recognition and generate clarinet sound as warning	Indoor, Outdoor	Internet Required	High
Mancini et al. [43]	Camera, PCB, and vibration motor	Not included	Obstacle recognition and vibration feedback for the direction	Outdoor	Offline	Low
Bauer et al. [44]	Camera, smartwatch, and smartphone	PASCAL VOC Dataset	Object detection with direction of object into audio output	Outdoor	Internet Required	High
Patil et al. [45]	Sensors, vibration motors,	No Dataset	Obstacle detection with audio output	Indoor, Outdoor	Offline	Low
Eckert et al. [46]	RGB-D camera and IMU sensors	PASCAL VOC dataset	Object detection with audio output	Indoor	Internet Required	High
Parikh et al. [47]	Smartphone, server, and headphone	Local dataset of 11 objects	Object detection with audio output	Outdoor	Internet Required	High
AL-Madani et al. [48]	BLE fingerprint, fuzzy logic	Not included	Localization of the person in the building	Indoor	Offline (Choice Wi-Fi or BLE)	Low
Proposed Method	RGB Camera, Distance Sensor, DSP processor, Headphone	Local dataset of highly relevant objects for VIP	Object detection, Count of objects, obstacle warnings, read text, and works in different modes	Indoor, Outdoor	Offline	Low

The performance of the proposed device has been tested on 36 people, including 20 visually impaired and 16 blind-folded people belonging to different age groups. The test is conducted for both indoor and outdoor environments. All were given sticks and a supporting person while using the proposed framework. Various rehabilitation workers and teachers working in this field were also involved to help to conduct the experiments smoothly. Before the trials, all those involved were briefly informed about the device so that users were aware about the experimentation steps. Different obstacles were used in an indoor environment, such as a chair, stairs, humans, walls, etc. While in outdoor environments, trained objects such as cars, humans, and vehicles were used. The visually impaired people previously used blind sticks, which give alerts for objects coming in front of the stick by means of vibrations. It takes a lot of mental effort and attention when walking only with the help of a stick. They experienced lots of problems while using the stick in crowed areas. After using this device, the proposed framework was found to be comfortable and easy to use in crowded areas. The developed technology is found to be highly useful, with which users can also understand the surrounding scenario easily while navigating without putting in too much effort. The proposed aid for visually impaired seems good in the sense that it does not need any prior knowledge about the position, shape and size of object and obstacles.

4. Conclusions and Future Scope

An assistive system is proposed for visually impaired persons through which they can perceive their surroundings and objects in real-time and navigate independently. Deep learning-based object detection, in assistance with various distance sensors, is used to make the user aware of obstacles, to provide safe navigation where all information is provided to the user in the form of audio. A list of highly relevant objects to visually impaired people is collected, and the dataset is prepared manually and is used to train the deep learning model for multiple epochs. Images are augmented and manually annotated to achieve more robustness. The results demonstrate 95.19% object detection accuracy and 99.69% object recognition accuracy in real-time. The proposed work uses 0.3 s for multi-instance and multi-object detection from the captured image, which is less than a non-visually impaired person in certain scenarios. The proposed assistive system gives more information with higher accuracy in

real time for visually challenged people. It can also easily differentiate between objects and obstacles coming in front of the camera.

Future work will focus on the inclusion of more objects in the dataset, which can make the dataset more efficient for the assistance of visually impaired people. More sensors will be associated with it to detect, for example, downstairs and other trajectories, giving a wider range of assistance to the visually impaired.

Author Contributions: Conceptualization, R.C.J., S.Y. and M.K.D., methodology, R.C.J., experimental setup, R.C.J., S.Y. and M.K.D.; writing-original draft preparation, R.C.J. and S.Y., writing—review and editing, M.K.D. and C.M.T.-G.; supervision, M.K.D. and C.M.T.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Grants from Department of Science and Technology, Government of India, grant number SEED/TIDE/2018/6/G.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Bourne, R.R.A.; Flaxman, S.R.; Braithwaite, T.; Cicinelli, M.V.; Das, A.; Jonas, J.B.; Keeffe, J.; Kempen, J.H.; Leasher, J.; Limburg, H.; et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: A systematic review and meta-analysis. *Lancet Glob. Health* **2017**, *5*, e888–e897. [CrossRef]
2. Global Trends in the Magnitude of Blindness and Visual Impairment. Available online: https://www.who.int/blindness/causes/trends/en/ (accessed on 12 June 2020).
3. Blindness and Vision Impairment. Available online: https://www.who.int/news-room/fact-sheets/detail/blindness-andvisual-impairment (accessed on 12 June 2020).
4. Thaler, L.; Arnott, S.R.; Goodale, M.A. Neural Correlates of Natural Human Echolocation in Early and Late Blind Echolocation Experts. *PLoS ONE* **2011**, *6*, e20162. [CrossRef] [PubMed]
5. Van Lam, P.; Fujimoto, Y.; Van Phi, L. A Robotic Cane for Balance Maintenance Assistance. *IEEE Trans. Ind. Inform.* **2019**, *15*, 3998–4009. [CrossRef]
6. Bhatlawande, S.; Mahadevappa, M.; Mukherjee, J.; Biswas, M.; Das, D.; Gupta, S. Design, Development, and Clinical Evaluation of the Electronic Mobility Cane for Vision Rehabilitation. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2014**, *22*, 1148–1159. [CrossRef] [PubMed]
7. Ando, B.; Baglio, S.; Marletta, V.; Valastro, A. A Haptic Solution to Assist Visually Impaired in Mobility Tasks. *IEEE Trans. Hum. Mach. Syst.* **2015**, *45*, 641–646. [CrossRef]
8. Bai, J.; Lian, S.; Liu, Z.; Wang, K.; Liu, D. Smart guiding glasses for visually impaired people in indoor environment. *IEEE Trans. Consum. Electron.* **2017**, *63*, 258–266. [CrossRef]
9. Di, P.; Hasegawa, Y.; Nakagawa, S.; Sekiyama, K.; Fukuda, T.; Huang, J.; Huang, Q. Fall Detection and Prevention Control Using Walking-Aid Cane Robot. *IEEE/ASME Trans. Mechatron.* **2015**, *21*, 625–637. [CrossRef]
10. López-De-Ipiña, D.; Lorido, T.; López, U. BlindShopping: Enabling Accessible Shopping for Visually Impaired People through Mobile Technologies. In Proceedings of the 9th International Conference on Smart Homes and Health Telematics, Montreal, QC, Canada, 20–22 June 2011; LNCS 6719. pp. 266–270.
11. Tekin, E.; Coughlan, J.M. An algorithm enabling blind users to find and read barcodes. In Proceedings of the 2009 Workshop on Applications of Computer Vision (WACV), Snowbird, UT, USA, 7–8 December 2009; pp. 1–8. [CrossRef]
12. Ahmad, N.S.; Boon, N.L.; Goh, P. Multi-Sensor Obstacle Detection System Via Model-Based State-Feedback Control in Smart Cane Design for the Visually Challenged. *IEEE Access* **2018**, *6*, 64182–64192. [CrossRef]
13. Takatori, N.; Nojima, K.; Matsumoto, M.; Yanashima, K.; Magatani, K. Development of voice navigation system for the visually impaired by using IC tags. In Proceedings of the 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, New York, NY, USA, 30 August–3 September 2006; pp. 5181–5184.

14. Fukasawa, A.J.; Magatani, K. A navigation system for the visually impaired an intelligent white cane. In Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, USA, 28 August–1 September 2012; pp. 4760–4763.
15. Rastogi, R.; Pawluk, T.V.D.; Ketchum, J.M. Intuitive Tactile Zooming for Graphics Accessed by Individuals Who are Blind and Visually Impaired. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2013**, *21*, 655–663. [CrossRef]
16. Aladrén, A.; Lopez-Nicolas, G.; Puig, L.; Guerrero, J.J. Navigation Assistance for the Visually Impaired Using RGB-D Sensor with Range Expansion. *IEEE Syst. J.* **2014**, *10*, 922–932. [CrossRef]
17. Yang, X.; Yuan, S.; Tian, Y. Assistive Clothing Pattern Recognition for Visually Impaired People. *IEEE Trans. Hum. Mach. Syst.* **2014**, *44*, 234–243. [CrossRef]
18. Katzschmann, R.K.; Araki, B.; Rus, D. Safe Local Navigation for Visually Impaired Users with a Time-of-Flight and Haptic Feedback Device. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 583–593. [CrossRef] [PubMed]
19. Kang, M.-C.; Chae, S.-H.; Sun, J.-Y.; Yoo, J.-W.; Ko, S.-J. A novel obstacle detection method based on deformable grid for the visually impaired. *IEEE Trans. Consum. Electron.* **2015**, *61*, 376–383. [CrossRef]
20. Chen, X.; Xu, J.; Yu, Z. A 68-mw 2.2 Tops/w Low Bit Width and Multiplierless DCNN Object Detection Processor for Visually Impaired People. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 3444–3453. [CrossRef]
21. Hassaballah, M.; Awad, A.I. *Deep Learning in Computer Vision*; Informa UK Limited: London, UK, 2020.
22. Mocanu, B.; Tapu, R.; Zaharia, T. DEEP-SEE FACE: A Mobile Face Recognition System Dedicated to Visually Impaired People. *IEEE Access* **2018**, *6*, 51975–51985. [CrossRef]
23. Ton, C.; Omar, A.; Szedenko, V.; Tran, V.H.; Aftab, A.; Perla, F.; Bernstein, M.J.; Yang, Y. LIDAR Assist Spatial Sensing for the Visually Impaired and Performance Analysis. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 1727–1734. [CrossRef]
24. Croce, D.; Giarre', L.; Pascucci, F.; Tinnirello, I.; Galioto, G.E.; Garlisi, D.; Valvo, A.L. An Indoor and Outdoor Navigation System for Visually Impaired People. *IEEE Access* **2019**, *7*, 170406–170418. [CrossRef]
25. Meshram, V.V.; Patil, K.; Meshram, V.A.; Shu, C. An Astute Assistive Device for Mobility and Object Recognition for Visually Impaired People. *IEEE Trans. Hum. Mach. Syst.* **2019**, *49*, 449–460. [CrossRef]
26. Goyal, S.; Bhavsar, S.; Patel, S.; Chattopadhyay, C.; Bhatnagar, G. SUGAMAN: Describing floor plans for visually impaired by annotation learning and proximity-based grammar. *IET Image Process.* **2019**, *13*, 2623–2635. [CrossRef]
27. Chang, W.-J.; Chen, L.-B.; Hsu, C.-H.; Chen, J.-H.; Yang, T.-C.; Lin, C.-P. MedGlasses: A Wearable Smart-Glasses-Based Drug Pill Recognition System Using Deep Learning for Visually Impaired Chronic Patients. *IEEE Access* **2020**, *8*, 17013–17024. [CrossRef]
28. Jarraya, S.K.; Al-Shehri, W.S.; Ali, M.S. Deep Multi-Layer Perceptron-Based Obstacle Classification Method from Partial Visual Information: Application to the Assistance of Visually Impaired People. *IEEE Access* **2020**, *8*, 26612–26622. [CrossRef]
29. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2009**, *88*, 303–338. [CrossRef]
30. Calik, R.C.; Demirci, M.F. Cifar-10 Image Classification with Convolutional Neural Networks for Embedded Systems. In Proceedings of the 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), Aqaba, Jordan, 28 October–1 November 2018; pp. 1–2.
31. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei, L.-F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
32. Xiao, J.; Hays, J.; Ehinger, K.A.; Oliva, A.; Torralba, A. SUN database: Large-scale scene recognition from abbey to zoo. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3485–3492.
33. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Bioinformatics Research and Applications*; Springer Science and Business Media LLC: Berlin, Germany, 2014; Volume 8693, pp. 740–755.
34. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
35. Redmon, J.; Farhadi, A. YOLO-v3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.

36. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
37. Qiu, X.; Yuan, C. Improving Object Detection with Convolutional Neural Network via Iterative Mechanism. *Bioinform. Res. Appl.* **2017**, *10636*, 141–150.
38. Günther, J.; Pilarski, P.M.; Helfrich, G.; Shen, H.; Diepold, K. First Steps towards an Intelligent Laser Welding Architecture Using Deep Neural Networks and Reinforcement Learning. *Procedia Technol.* **2014**, *15*, 474–483. [CrossRef]
39. Guerra, E.; De Lara, J.; Malizia, A.; Díaz, P. Supporting user-oriented analysis for multi-view domain-specific visual languages. *Inf. Softw. Technol.* **2009**, *51*, 769–784. [CrossRef]
40. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1646–1654.
41. Hoang, V.-N.; Nguyen, T.-H.; Le, T.-L.; Tran, T.-H.; Vuong, T.-P.; Vuillerme, N. Obstacle detection and warning system for visually impaired people based on electrode matrix and mobile Kinect. *Vietnam. J. Comput. Sci.* **2016**, *4*, 71–83. [CrossRef]
42. Yang, K.; Bergasa, L.M.; Romera, E.; Cheng, R.; Chen, T.; Wang, K. Unifying terrain awareness through real-time semantic segmentation. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 1033–1038. [CrossRef]
43. Mancini, A.; Frontoni, E.; Zingaretti, P. Mechatronic System to Help Visually Impaired Users during Walking and Running. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 649–660. [CrossRef]
44. Bauer, Z.; Dominguez, A.; Cruz, E.; Gomez-Donoso, F.; Orts-Escolano, S.; Cazorla, M. Enhancing perception for the visually impaired with deep learning techniques and low-cost wearable sensors. *Pattern Recognit. Lett.* **2019**. [CrossRef]
45. Patil, K.; Jawadwala, Q.; Shu, F.C. Design and Construction of Electronic Aid for Visually Impaired People. *IEEE Trans. Hum. Mach. Syst.* **2018**, *48*, 172–182. [CrossRef]
46. Eckert, M.; Blex, M.; Friedrich, C.M. Object Detection Featuring 3D Audio Localization for Microsoft HoloLens—A Deep Learning based Sensor Substitution Approach for the Blind. In Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies, Funchal, Portugal, 19–21 January 2018; pp. 555–561.
47. Parikh, N.; Shah, I.; Vahora, S. Android Smartphone Based Visual Object Recognition for Visually Impaired Using Deep Learning. In Proceedings of the 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 3–5 April 2018; pp. 420–425.
48. Al-Madani, B.; Orujov, F.; Maskeliunas, R.; Damaševičius, R.; Venckauskas, A. Fuzzy Logic Type-2 Based Wireless Indoor Localization System for Navigation of Visually Impaired People in Buildings. *Sensors* **2019**, *19*, 2114. [CrossRef] [PubMed]

International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 03 - Issue 06 || June 2017 || PP. 09-15

Image to Speech Conversion for Visually Impaired

Asha G. Hagargund[1], Sharsha Vanria Thota[2], Mitadru Bera[3], Eram Fatima Shaik[4]

[1]*Assistant Professor, Department of Electronics and Communication Engineering, BMSIT&M, Bangalore*
Affiliated to Visvesvaraya Technological University, Belgaum, India
[2]*Dept. of Electronics and Communication Engineering, BMSIT&M, Bengaluru*
Affiliated to Visvesvaraya Technological University, Belgaum, India
[3]*Dept. of Electronics and Communication Engineering, BMSIT&M, Bengaluru*
Affiliated to Visvesvaraya Technological University, Belgaum, India
[4]*Dept. of Electronics and Communication Engineering, BMSIT&M, Bengaluru*
Affiliated to Visvesvaraya Technological University, Belgaum, India

Abstract: Visual impairment is one of the biggest limitation for humanity, especially in this day and age when information is communicated a lot by text messages (electronic and paper based) rather than voice. The device we have proposed aims to help people with visual impairment. In this project, we developed a device that converts an image's text to speech. The basic framework is an embedded system that captures an image, extracts only the region of interest (i.e. region of the image that contains text) and converts that text to speech. It is implemented using a Raspberry Pi and a Raspberry Pi camera. The captured image undergoes a series of image pre-processing steps to locate only that part of the image that contains the text and removes the background. Two tools are used convert the new image (which contains only the text) to speech. They are OCR (Optical Character Recognition) software and TTS (Text-to-Speech) engines. The audio output is heard through the raspberry pi's audio jack using speakers or earphones.

Keywords: Embedded system, OCR, pre-processing, Raspberry Pi, TTS

1. Introduction

In our planet of 7.4 billion humans, 285 million are visually impaired out of whom 39 million people are completely blind, i.e. have no vision at all, and 246 million have mild or severe visual impairment (WHO, 2011). It has been predicted that by the year 2020, these numbers will rise to 75 million blind and 200 million people with visual impairment [7]. As reading is of prime importance in the daily routine (text being present everywhere from newspapers, commercial products, sign-boards, digital screens etc.) of mankind, visually impaired people face a lot of difficulties. Our device assists the visually impaired by reading out the text to them.

There have been numerous advances in this area to help visually impaired to read without much difficulties. The existing technologies use a similar approach as mentioned in this paper, but they have certain drawbacks. Firstly, the input images taken in previous works have no complex background, i.e. the test inputs are printed on a plain white sheet. It is easy to convert such images to text without pre-processing, but such an approach will not be useful in a real-time system [1][2][3]. Also, in methods that use segmentation of characters for recognition, the characters will be read out as individual letter and not a complete word. This gives an undesirable audio output to the user. For our project, we wanted the device to be able to detect the text from any complex background and read it efficiently. Inspired by the methodology used by Apps such as "CamScanner", we assumed that in any complex background, the text will most likely be enclosed in a box eg billboards, screens etc. By being able to detect a region enclosing four points, we assume that this is the required region containing the text. This is done using warping and cropping. The new image obtained then undergoes edge detection and a boundary is then drawn over the letters. This gives it more definition. The image is then processed by the OCR and TTS to give audio ouput.

1.1. BASIC BLOCK DIAGRAM

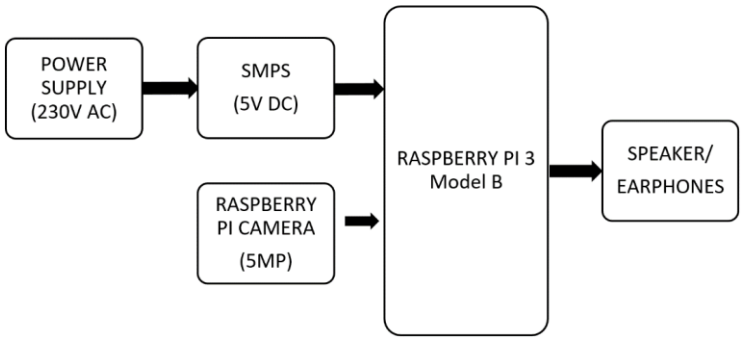


Fig 1.

The device consists of a Raspberry Pi 3B, speaker or earphones, Raspberry pi camera, power supply (230V AC) and a switched mode power supply (SMPS). The SMPS converts the 230V AC power supply to 5V DC to power the Raspberry Pi. The camera must manually be pointed towards the text and a picture is captured. This picture is then processed by the Raspberry Pi and the audio output is heard through the speaker.

1.2. OCR ENGINE

The extraction of the text in the image is done using optical character recognition (OCR). OCR is a field of research in pattern recognition, artificial intelligence and computer vision. It is the conversion of the images of typed, handwritten or printed text into a digital text or computer format text. Earlier OCR versions had to be trained in each character of a text with its specific font. Today, advanced OCRs are available that have a high degree of accuracy, support a wide variety of image formats, languages and fonts. For our project, we have used Tesseract OCR. It is the most accurate open source OCR engine and is powered by google. It can be used on the Linux, mac and windows platform. The newest Tesseract version, 3.4 supports a hundred languages. However, images must undergo a number of pre-processing stages like noise removal, scaling etc. otherwise the output will be of low quality.

1.3. TTS SOFTWARE

The process of converting text to speech by a computer is called speech synthesis. A text to speech system(TTS) is used to perform speech synthesis. A TTS is composed of two parts: front end and back end. The front end converts the text to a symbol, for example, a number. Each symbol generated is assigned a phonetic. The back end then converts the phonetic into sound. In our project, we have used Festival TTS. Festival is the most widely used open source TTS. It has a wide variety of voices and support English, Spanish and welsh language. We have used the English language.

2. Motivation

Our device is designed for people with mild or moderate visual impairment by providing the capability to listen to the text. It can also act as a learning aid for people suffering from dyslexia or other learning disabilities that involve difficulty in reading or interpreting words and letters. We wish to enable these people to be independent and self-reliant as they will no longer need assistance to understand printed text. Such people will always have access to information hence they will never feel at a disadvantage. The impact of the development and introduction of our system into the technological world will be a revolutionary boon to modern civilization.

3. Literature Survey

Visual impairment or vision loss is defined as the decreased ability to see clearly and cannot be fixed using glasses. Blindness is the term used for complete vision loss. The common causes of vision loss are uncorrected refractive errors, cataracts and glaucoma. People with visual impairment face a number of difficulties in normal daily activities like walking, driving and reading.[9]

3.1. BRAILLE

Braille is writing and reading system used people who have visual impairment. Braille language is written on embossed paper. The braille characters are small rectangular blocks called cells that contain bumps called raised dots. The visually impaired person feels the arrangement of the raised dots which conveys the information. [10]

Braille literacy statistics of India: One out of every three blind people in the world is an Indian. It is estimated that nearly 15 million Indians are blind and out of that 2 million are children. Only 5% of the children receive education. Although braille readers, keyboards and monitors exist, they are not accessible to the rural communities and braille material is not easily and abundantly available. [11]

3.2. RASPBERRY PI

The raspberry Pi is a small, low cost CPU which can be used with a monitor, keyboard and mouse to become an efficient, full-fledged computer [12]. The reason we chose Raspberry Pi micro-computer for our project is that, firstly, it is an easily available, low-cost device. RPi uses software which are either free or open source, which also makes it cost-effective. The Raspberry Pi uses an SD card for storage and its small size also gives us the advantages of portability.
[13]

As a part of the software development, the Open CV (Open source Computer Vision) libraries are utilized for image processing. Each function and data structure was designed with the Image Processing coder in mind. [14]

3.3. EXISTING SYSTEMS AND THEIR LIMITATIONS

- One of the biggest advantages of barcode readers is portability. Hence, they can be used by the visually impaired in identifying different products. An extensive database is created which contains all the information about the product. The user simply scans the bar code and the product details are listed through e-braille readers. The disadvantage with this product is that the user might not be able to point the bar code reader in the correct direction. [2]
- Another approach is optical enhancement solutions such as an optical zooming device that expands the braille character. However, not all visually impaired people need to know braille language. [4]
- Some methods aim at converting text to speech. This is accomplished using a scanner, speakers and a computer. This method is efficient only with simple scanned documents. It cannot extract text from an image with a complex background. [4]

4. System Specifications

4.1.1. SOFTWARE SPECIFICATIONS

Raspbian is a free operating system, based on Debian, optimized for the Raspberry Pi hardware. Raspbian Jessie is used as the version is RPi's main operating system in our project. Our code is written in Python language (version 2.7.13) and the functions are called from OpenCV. OpenCV, which stands for Open Source Computer Vision, is a library of functions that are used for real-time applications like image processing, and many others [14]. Currently, OpenCV supports a wide variety of programming languages like C++, Python, Java etc. and is available on different platforms including Windows, Linux, OS X, Android, iOS etc. [15]. The version used for our project is opencv-3.0.0. OpenCV's application areas include Facial recognition system, Gesture recognition, Human–computer interaction (HCI), Mobile robotics, Motion understanding, Object identification, Segmentation and recognition, Motion tracking, Augmented reality and many more. For performing OCR and TTS operations we install Tesseract OCR and Festival software. Tesseract is an open source Optical Character Recognition (OCR) Engine, available under the Apache 2.0 license. It can be used directly, or (for programmers) using an API to extract typed, handwritten or printed text from images. It supports a wide variety of languages. The package is generally called 'tesseract' or 'tesseract-ocr'.

Festival TTS was developed by the "The Centre for Speech Technology Research",UK. It is an open source software that has a framework for building efficient speech synthesis systems. It is multi-lingual (supports British English, American English and Spanish). As Festival is a part of the package manager for Raspberry Pi, it is easy to install.

4.1.2. HARDWARE SPECIFICATIONS

Raspberry pi is a device that contains several important functions on a single chip. It is a system on a chip(SoC). The Raspberry Pi 3 uses Broadcom BCM2837 SoC Multimedia processor. The Raspberry Pi's CPU is the 4x ARM Cortex-A53, 1.2GHz processor. It has internal memory 1GB LPDDR RAM (900Mhz) and external memory can be extended to 64 GB. In Raspberry Pi 3, the two main new features are wireless internet connection 802.11n and Bluetooth 4.1 classic. It has 40 GPIO pins. [16] The Raspberry pi camera is 5MP and

has a resolution of 2592x1944. The Raspberry Pi has a 3.5mm audio port so earphones or speaker can easily be connected to it to hear audio.

5. Methodology

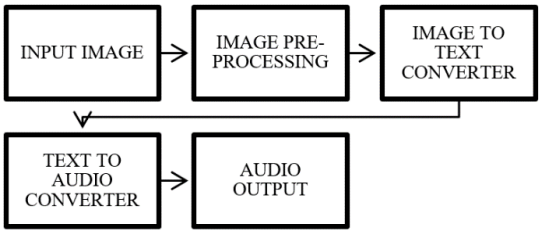


Fig 2.

Image acquisition: In this step, the inbuilt camera captures the images of the text. The quality of the image captured depends on the camera used. We are using the Raspberry Pi's camera which 5MP camera with a resolution of 2592x1944.

Image pre-processing: This step consists of color to gray scale conversion, edge detection, noise removal, warping and cropping and thresholding. The image is converted to gray scale as many OpenCV functions require the input parameter as a gray scale image. Noise removal is done using bilateral filter. Canny edge detection is performed on the gray scale image for better detection of the contours. The warping and cropping of the image are performed according to the contours. This enables us to detect and extract only that region which contains text and removes the unwanted background. In the end, Thresholding is done so that the image looks like a scanned document. This is done to allow the OCR to efficiently convert the image to text.

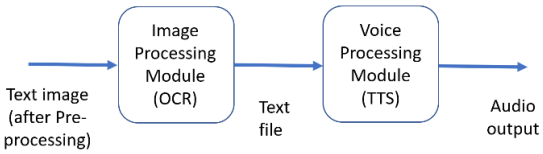


Fig 3.

Image to text conversion: The above diagram(fig.3) shows the flow of Text-To-Speech. The first block is the image pre-processing modules and the OCR. It converts the pre-processed image, which is in .png form, to a .txt file. We are using the Tesseract OCR.

Text to speech conversion: The second block is the voice processing module. It converts the .txt file to an audio output. Here, the text is converted to speech using a speech synthesizer called Festival TTS. The Raspberry Pi has an on-board audio jack, the on-board audio is generated by a PWM output.

6. Results

The obtained output images after pre-processing are displayed below. Figure 4 shows the original image that was captured using the Pi Camera. Figure 5 to Figure 11 display the pre-processing done in each stage. And finally Figure 11 represents the image which is given as input to the OCR. Figure 12 displays the text obtained at the output of the OCR engine. It is evident that the result is not completely accurate. This is because of the less resolution of the camera used. Better results can be obtained if the camera used is a High definition camera.

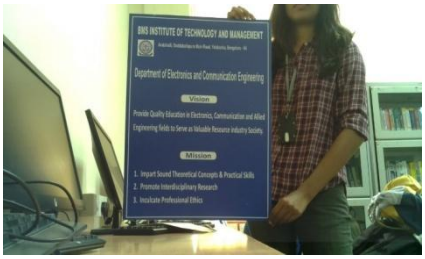


Fig 4: Original image captured from the camera

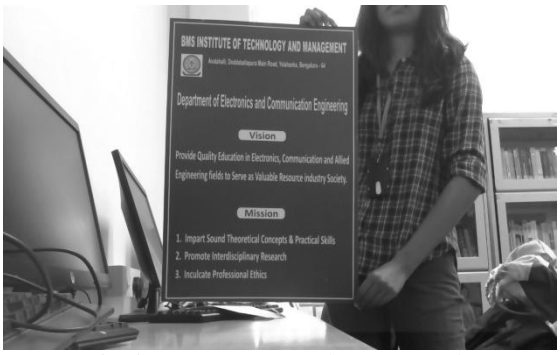


Fig 5: Image converted to gray scale

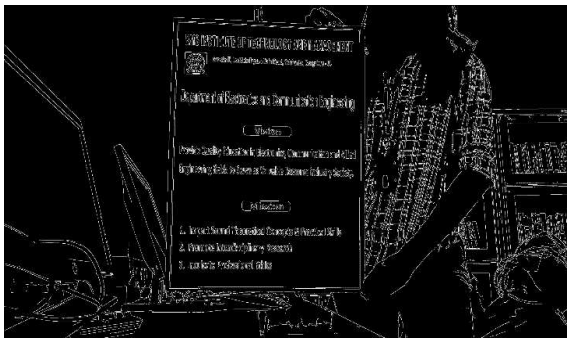


Fig 6: Performing edge detection

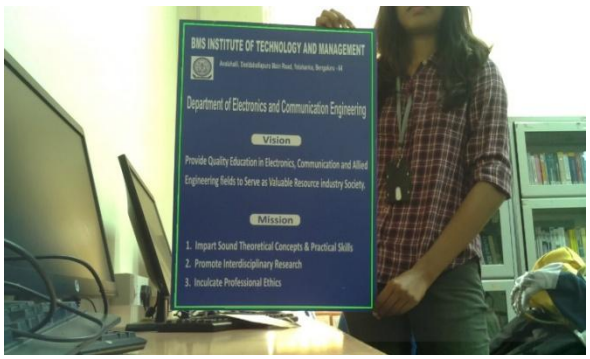


Fig 7: Contour detection

Fig 8: Warped and cropped image

Fig 9: Sharpening the image

Fig 10: Convert to grayscale before thresholding

Fig 11: Thresholding

```
BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT
(ï¬rst Aualahalii, Doddahallapura Main Road, Yelahanlca.
Bengaluru - 64

Department of Electronics and Communication Engineering
Provide Quality Education in Electronics, Communication and
Allied
Engineering fields to Serve as Valuable Resource industry Society.
1. Import Sound Theoretical Concepts & Practical Skills
2. Promote Interdisciplinary Research
3. Inculcate Professional Ethics
```

Fig 12: Tesseract output

7. Conclusion

The system enables the visually impaired to not feel at a disadvantage when it comes to reading text not written in braille. The image pre-processing part allows for the extraction of the required text region from the complex background and to give a good quality input to the OCR. The text, which is the output of the OCR is sent to the TTS engine which produces the speech output. To allow for portability of the device, a battery may be used to power up the system. The future work can be developing devices that perform object detection and extracting text from videos instead of static images.

References

[1]. D.Velmurugan, M.S.Sonam, S.Umamaheswari, S.Parthasarathy, K.R.Arun[2016]. *A Smart Reader for Visually Impaired People Using Raspberry PI*. International Journal of Engineering Science and Computing IJESC Volume 6 Issue No. 3.

[2]. K Nirmala Kumari, Meghana Reddy J [2016]. *Image Text to Speech Conversion Using OCR Technique in Raspberry Pi.* International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 5, Issue 5, May 2016.

[3]. Silvio Ferreira, C´eline Thillou, Bernard Gosselin. *From Picture to Speech: An Innovative Application for Embedded Environment*. Faculté Polytechnique de Mons, Laboratoire de Théorie des Circuits et Traitement du Signal Bˆatiment Multitel - Initialis, 1, avenue Copernic, 7000, Mons, Belgium.

[4]. Nagaraja L, Nagarjun R S, Nishanth M Anand, Nithin D, Veena S Murthy [2015]. *Vision based Text Recognition using Raspberry Pi.* International Journal of Computer Applications (0975 – 8887) National Conference on Power Systems & Industrial Automation.

[5]. Poonam S. Shetake, S. A. Patil, P. M. Jadhav [2014] *Review of text to speech conversion methods.*

[6]. International Journal of Industrial Electronics and Electrical Engineering, ISSN: 2347-6982 Volume-2, Issue-8, Aug.-2014.

[7]. S. Venkateswarlu, D. B. K. Kamesh, J. K. R. Sastry, Radhika Rani [2016] *Text to Speech Conversion.* Indian Journal of Science and Technology, Vol 9(38), DOI: 10.17485/ijst/2016/v9i38/102967, October 2016.

[8]. World Health Organization. 10 facts about blindness and visual impairment. 2015. Available from: http://www.who. int/features/factfiles/blindness/blindness_facts/en/

[9]. http://elinux.org/RPi_Text_to_Speech_(Speech_Synthesis)

[10]. https://en.wikipedia.org/wiki/Visual_impairment

[11]. https://en.wikipedia.org/wiki/Braille

[12]. https://www.classycyborgs.org/braille-literacy-statistics-india/

[13]. www.raspberrypi.org

[14]. http://www.zdnet.com/article/raspberry-pi-11-reasons-why-its-the-perfect-small-server/

[15]. http://aishack.in/tutorials/opencv/

[16]. http://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_setup/py_intro/py_intro.html

[17]. http://hackaday.com/2016/02/28/introducing-the-raspberry-pi-3/

Intelligent Interfaces for Assisting Blind People using Object Recognition Methods

Article · May 2022

CITATIONS
2

READS
83

5 authors, including:

Jamil Alsayaydeh
Technical University of Malaysia Malacca
75 PUBLICATIONS **265** CITATIONS
SEE PROFILE

Maslan Zainon
Technical University of Malaysia Malacca
30 PUBLICATIONS **78** CITATIONS
SEE PROFILE

Intelligent Interfaces for Assisting Blind People using Object Recognition Methods

Jamil Abedalrahim Jamil Alsyayadeh[1]*
Hasvinii Baskaran[4]
Department of Electronics & Computer Engineering Technology, Fakulti Teknologi Kejuruteraan Elektrik & Elektronik (FTKEE), Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia

Irianto[2]
Department-General Education, Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates

Maslan Zainon[3]
Department of Electrical Engineering Technology
Fakulti Teknologi Kejuruteraan Elektrik & Elektronik (FTKEE), Universiti Teknikal Malaysia Melaka (UTeM)
Melaka, Malaysia

Safarudin Gazali Herawan[5]
Industrial Engineering Department
Faculty of Engineering, Bina Nusantara University
Jakarta, Indonesia 11480

***Abstract*—Object recognition method is a computer vision technique for identifying objects in images. The main purpose of this system build is to put an end to blindness by constructing automated hardware with Raspberry Pi that enables a visually impaired person to detect objects or persons in front of them instantly, and inform what is in front of them through audio. Raspberry Pi receives data from a camera then processes it. In addition, the blind will listen to a voice narration via an audio receiver. This paper's key objective is to provide the blind with cost-effective smart assistance to explore and sense the world independently. The second objective is to provide a convenient portable device allows users to recognise objects without touch, having the system determine the object in front of them. The camera module attached in Raspberry Pi will capture image and the processor will then process it. Subsequently, the processed image sends data to the audio receiver narrating the detected object(s). This system will be very useful for a blind person to explore the world by listening to the voice narration. The generated voice narration after processing the image will help the blind to visualise objects in front of them.**

Keywords*—Object recognition method; computer vision; blind people; image processing; Raspberry Pi; Pi camera; smart assistance; portable device; voice narration; visualise

I. Introduction

Physical movement is a challenge for the blind. Visual disability stands out from the many extreme obstacles affecting a person. This is clearly explained in this research study about the visually impaired patients. They face physical and social constraints accrued from their visual loss, and they need to improve on their health and independence [1][2][3]. Designing a gadget to help the blind is not something new. Various technologies exist to help the visually impaired, and as innovations increasingly propelled, ideas appear to provide intriguing measures to help the visually impaired. In any case, designing a device to aid blindness comes with a price and does not come easy, as it is often regarded as a so-called luxurious item in most developing nations.

As indicated by the World Health Organization (WHO) research study, [4][5] it has been estimated that over 1.3 billion people around the world have some form of vision impairment. Roughly, 80% of all kinds of vision debilitation are viewed as avoidable. Additionally, the WHO also mentioned that most visually impaired adolescents would require visual recovery intercessions for self-improvement. Nevertheless, it is most often that visual recovery treatments come with substantial hospital expenses, and with 90% of disabled people living with financial difficulties, visual restoration is not the best alternative for all. For full psychological improvement and better independence without bearing costly bills, blind people require an assistive device that helps them with their daily activities. There exist many assistive devices for visually impaired people and it became the inspiration in the background of this research study. Smart assistance such as a smart and autonomous walking stick, smart glasses/spectacles, or prosthetics [6][7][8][9]. The assistance from another individual is not always accessible, and are unfavoured by visually impaired individuals that search for freedom, without having to bear the cost of such expensive smart assistance equipment as well. We propose an audio receiver for blind people that uses real-time smart assistance interfaces and object recognition technologies as a solution to this occurring issue. Our system mainly consists of two components, Raspberry Pi and Pi camera. The smart assistance audio guidance was developed to assist users to determine objects in front of them and help them visualise the environment around them. The camera in the processor will capture image, then processes that image, and a voice narration will be sent through audio receiver.

The main objective is to provide the blind cost-effective smart assistance to explore and feel the world independently. This enables the blind to visualise their surroundings and afford current technologies. Additionally, we also aim to provide a portable device which is easy to utilise and permits them to recognise objects without touching, and describes the surroundings in front of them.

*Corresponding Author.

This system is used to assist blind people with voice narration, processed by the Raspberry Pi processor. This portable electronic device's purpose is to give voice narration informing what is in front of them. An important objective is to provide a portable device that is simple to use and low cost and affordable smart assistance to blind people. Another goal is to extend the computerised electronic travel aid for the blind by applying real-time object recognition technology. This blind guidance system is solid and financially perceptive. Real-time based smart assistance interfaces the audio receiver for blind people with voice narration by using object recognition methods to provide the blind with cost-effective smart assistance to explore and sense the world independently. Audio guidance helps them to know what is happening around them and it helps them to visualise their surroundings. By using real-time, the system will recognise the objects faster.

The remaining of this paper has been organized as follows: Section 2 discusses the related works. The background of the study is described in Section 3. Section 4 described the system implementation and testing. Section 5 described the results and discussion and finally, the conclusion is described in Section 6.

II. Related Work

There are a lot of assistive devices for visually impaired people to sense the world independently. All these devices rely mainly on ultrasonic sensors and Brailling.

A. EyeCane and EyeMusic

Maidenbaum et al. [10] designed EyeCane and EyeMusic to improve upon, or likely be within the far distant future, to update the traditional white cane. By applying statistics at visually far distances (5 meters) and greater angles, and most significantly by means of discarding contacts among the cane, and the user's surroundings in cluttered or indoor environments. The EyeCane converts point-distance information into aural and tactile signals. The Prototype of EyeCane and EyeMusic is shown in Fig. 1. The tool can provide distance information to the customer from two different directions at the same time: immediately in advance for long-distance perception and detection of waist-height obstacles, and pointing downward at a 45° angle for ground-level evaluation.

Fig. 1. Prototype of EyeCane and EyeMusic.

B. Blitab

Blitab is a device nicknamed "the iPad for the visually impaired". It appears similar to a digital book, however, its screen utilises smart liquids that protrude tactile pixels to show braille letters, making it conceivable for the blind to see entire pages of braille message at once. Perkins-style keyboard application, text-to-speech yield, and touch navigation provide a completely new user experience for braille and non-braille blind individuals. It empowers the fast conversion of any content into braille. Blitab is a platform for all current and future programming applications for visually impaired people, it is not only a tablet. The Prototype of Blitab is shown in Fig. 2.

Blitab is the world's first real tactile tablet designed specifically for the blind and visually impaired. The device's revolutionary smart liquid technology also allows it to display material images for blind people who do not use braille [11].

C. BrainPort V100

According to Grant et al. [12], BrainPort V100 is an oral electronic vision aid that uses electro-tactile stimulation to help profoundly blind people with direction, mobility, and object recognition. The device is used in conjunction with other assistive devices like a normal white cane or a guide dog.

It deciphers digital data from a wearable camcorder into delicate electrical incitement designs on the outside of the tongue. Users feel moving bubble-like patterns on their tongue then they figure out how to interpret or visualise according to the shape, size, area, and movement of articles in their condition. A few clients have portrayed it as having the option to "see with your tongue". What makes it extraordinary is seeing with your mouth may appear to be outlandish at first, yet with at least 10 hours of one-on-one instructional courses, wearers can figure out how to comprehend the shivers and "see" where objects are found, yet additionally, their size, shape and in the event that they are moving. In a clinical preliminary, 69% of members had the option that effectively recognises protests in an acknowledgment test following one year of preparing with the BrainPort. The Prototype of BrainPort V100 is shown in Fig. 3.

Fig. 2. Prototype of Blitab.

Fig. 3. Prototype of BrainPort V100.

TABLE I. COMPARISON BETWEEN EXISTING SYSTEM

System	Devices needed	Cost	accessibility	purpose
BrainPort V100	Headset, Intra Oral Device(IOD)	Expensive ($10000)	Controller	To provide oral electronic vision aid
EyeCane& EyeMusic	Infrared emitters, web camera, smartphone, headset	Low cost	Infrared sensors	To provide navigation control and identifies colour, shape and location of objects.
Blitab	Touch screen tablet, Braille lines	Low cost($500)	Braille display	Displays tactile images to blind people.
Proposed method	Raspberry Pi 3, Pi camera module, Headset	Cheap ($100)	Text-to-Speech module	The generated narration will be the final output that is transmitted to the user through a headset.

The difference between existing systems is shown in Table I.

III. BACKGROUND OF THE STUDY

Object Recognition is a method used in image processing to recognise real objects. This method is clearly explained in a research study about the importance of process that will help blind people to identify their daily items that are commonly used. Our system provides some kind of visual aid that recognises objects dynamically [13]. The algorithm used in this system analyses the object. For instance, a blind person is sitting on his dining table. He has multiple objects in front of him such as bottle, chair, dining table, etc. Therefore, our system will help him by narrating what is in front of them. Text-to-Speech module is used to convert text to speech. The text that is written in text file is the output of object detection. Google API is used for conversion of Text-to-Speech dynamically, provided that the internet connection is stable. This has been studied from a research that explains about Google API that is used for text-to-speech [14]. For example, if the camera captures a book in front of it, it detects the book and converts it into text from the image captured. The text will be written in a text file and then converted to speech by using Google Text-to-Speech. The architecture of this proposed system is the Raspberry Pi board. Raspberry Pi controller controls the system and activates the output and sends the instructions. The detailed specifications of Raspberry Pi 3 B+ consists of: four USB ports, an Ethernet port, forty GPIO pins, SD card slot, SOC (system on a chip), a DSI display interface, HDMI port, LAN controller, audio jack, CSI camera interface, RCA video socket, and 5V micro USB connector [15].

The Block diagram of the object recognition process is shown in Fig. 4.

The Pi camera is connected to a CSI camera interface of Raspberry Pi processor. The processor has an operating system named Raspbian, which process the image, voice narration and other conversions. The headset will connect to an audio jack for audio output. Once the system components activate, the camera module will begin a video stream of its front view, and the image in video will be processed. Before this process starts, the Raspberry Pi will create a video frame, activates "cv" environment, and runs the python script to activate the system. Thereafter, the processed image undergoes object detection for image classification and recognition. Hence, the image in the video will detect through real-time object recognition, and the label of each object will be printed in a text file, which is used for voice narration. The labels in the text file use Google Text-to-Speech for voice narration. The generated narration will be the final output that is transmitted to the user through a headset.

The flowchart of the object recognition process is shown in Fig. 5.

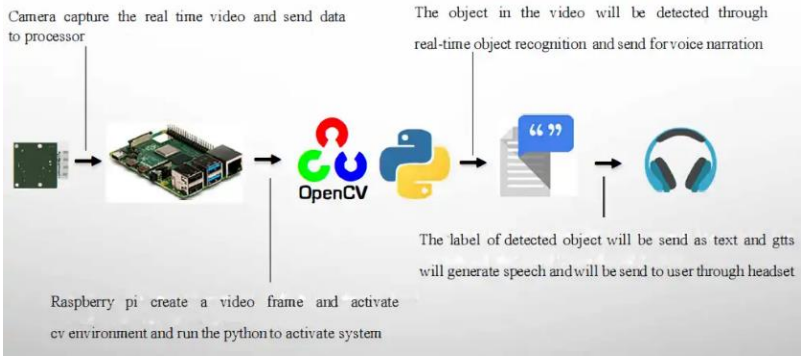


Fig. 4. Block Diagram of the Object Rprocess.

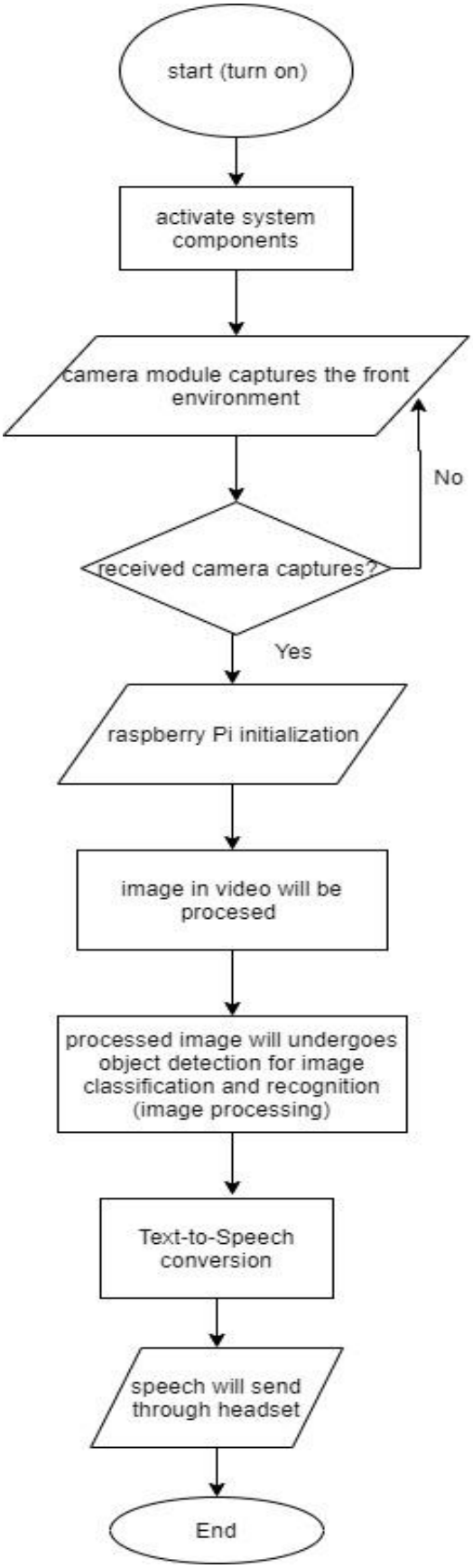


Fig. 5. Flowchart of the Object Recognition Process.

IV. System Implementation and Testing

A. Hardware Implementation

The necessary components in developing this system consist of a Raspberry Pi and Pi camera. The New Out of Box Software (Noobs) is installed on an SD card to format the Raspberry Pi that will be fixed in the Raspberry Pi, as studied in the manual that was given to study about Raspberry Pi startup [16]. Noobs contain Java SE Platform Products. It is an operating system installer with Raspbian pre-loaded. Once done, this Raspberry Pi will connect to power, start to boot, and be ready to use the operating system, whereas the Pi camera will be configured beforehand. The camera's interfacing option in Raspbian OS will be enabled manually to allow the camera to work with the system. Once the configuration is done, the Raspbian enables the camera. The image captured after configuring the Pi camera is shown in Fig. 6.

B. Software Implementation

The Raspbian operating system is used in the Raspberry Pi 3 model B+ as a platform to run this system; which is, the platform to create, run, and troubleshoot the coding of the software that has been used. Python IDLE software was used to build this system. Python IDLE ran in an OpenCV environment. OpenCV was created to provide a common infrastructure for computer vision applications such as deep learning, optical character recognition (OCR) and object detection, and more as explained in the article [17][18]. OpenCV-Python is the Python API for OpenCV. It's a Python bindings library aimed in solving computer vision challenges. Python has been enhanced with C/C++, enabling programmers to write/express code and develop Python wrappers that can be used as Python modules, as stated in the article named Python. It is packaged as an optional part of the Python packaging with many Linux distributions [19][20]. The actual code will run in the background of the CV environment. To write the necessary codes to run the system, the Python 3.7.3 was used. To capture the image, a Pi camera connected to a Raspberry Pi was used. Furthermore, code will be used to initialise the captured image.

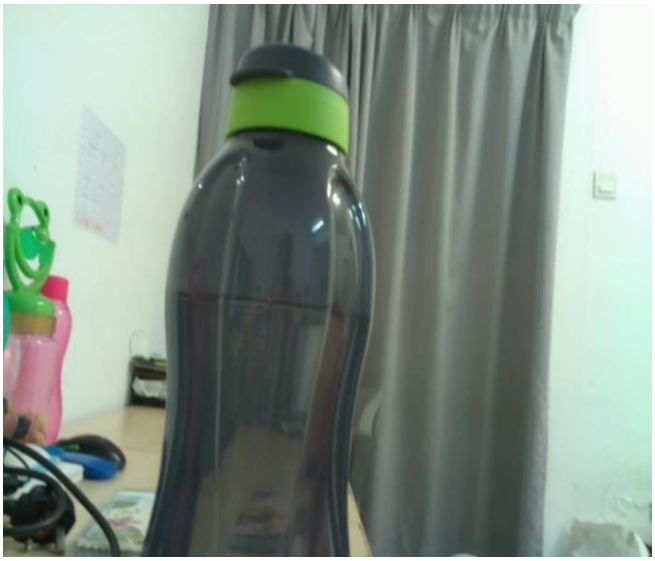


Fig. 6. Image Captured after Configuring the Pi Camera.

V. Result and Discussion

Here is a sample of coding and result for this proposed system. Some discussions are added up as an explanation to understand its function clearly.

A. Coding

Partially applied programming codes are displayed below in Fig. 7 to Fig. 10.

```
ap = argparse.ArgumentParser()
ap.add_argument("-p", "--prototxt", required=True,
    help="path to Caffe 'deploy' prototxt file")
ap.add_argument("-m", "--model", required=True,
    help="path to Caffe pre-trained model")
ap.add_argument("-c", "--confidence", type=float, default=0.2,
    help="minimum probability to filter weak detections")
args = vars(ap.parse_args())
```

Fig. 7. Construction of the Argument Parse.

The preceding code demonstrates how to construct an argument parse to parse the arguments. 'ArgumentParser()' converts the argument value from a string into some other type. The first line sets up an argument parser, followed by three mandatory command-line arguments. Firstly, the 'prototxt' is the path to the Caffe prototxt file which is known as the solver.prototxt, secondly, a configuration file, whereas '—model' is the path to the pre-trained model and thirdly, the '—confidence' is minimum probability threshold when filtering weak detections and it is set to 20% by default.

```
CLASSES = ["background", "aeroplane", "bicycle", "bird", "boat",
    "bottle", "bus", "car", "cat", "chair", "cow", "diningtable",
    "dog", "horse", "motorbike", "person", "pottedplant", "sheep",
    "sofa", "train", "tvmonitor"]
COLORS = np.random.uniform(0, 255, size=(len(CLASSES), 3))

# load our serialized model from disk
print("[INFO] loading model...")
net = cv2.dnn.readNetFromCaffe(args["prototxt"], args["model"])
```

Fig. 8. Initialisation of 'Classes'.

These lines of code initialise 'CLASSES', class labels, and equivalent COLORS, for on-frame text and bounding boxes. Furthermore, the last line loads the serialised neural network model.

```
for i in np.arange(0, detections.shape[2]):
    # extract the confidence (i.e., probability) associated with
    # the prediction
    confidence = detections[0, 0, i, 2]

    # filter out weak detections by ensuring the `confidence` is
    # greater than the minimum confidence
    if confidence > args["confidence"]:
        # extract the index of the class label from the
        # `detections`, then compute the (x, y)-coordinates of
        # the bounding box for the object
        idx = int(detections[0, 0, i, 1])
        box = detections[0, 0, i, 3:7] * np.array([w, h, w, h])
        (startX, startY, endX, endY) = box.astype("int")

        # draw the prediction on the frame
        label = "{}: {:.2f}%".format(CLASSES[idx],
            confidence * 100)
        cv2.rectangle(frame, (startX, startY), (endX, endY),
            COLORS[idx], 2)
        y = startY - 15 if startY - 15 > 15 else startY + 15
        cv2.putText(frame, label, (startX, y),
            cv2.FONT_HERSHEY_SIMPLEX, 0.5, COLORS[idx], 2)
```

Fig. 9. Looping the Detection.

This part explains how this system is able to detect numerous objects in a single image. First step is to loop over the detections. The chance of each detection will be checked and tallied with confidence. If the confidence exceeds the threshold, the prediction will be displayed in terminal and drawn on the frame. Detections will undergo loops and its confidence value is extracted in each loop. Therefore, the class label index is extracted if the confidence level is greater than the minimal threshold, as well as the bounding box coordinates surrounding the detected objects that have been computed too, and a rectangle displaying text is created on the detected object. Labels containing CLASS name and confidence build, and displayed as the processed-colored rectangle created around the object. Finally, the system computes the colored text that was generated onto the frame by using the y-value.

```
from gtts import gTTS
from time import sleep
import os
import pyglet

file = open("label.txt", "r").read().replace("\n", " ")
language = 'en'
tts = gTTS(text = 'hi there. i am here to assist you to look the world in front of you. there is' + str(file) + 'in front of you' , lang = 'en', slow = False
tts.save('voice.mp3')
os.system("start /home/pi/Desktop/real-time-object-detection/voice.mp3")
print(file)
```

Fig. 10. Voice Narration.

To enable the Raspberry PI to "talk", the Google Text to Speech (gTTS) module is used in Python is used and also imported into the Raspbian system. This is used to command the system to read the image classification result that has been written after the real time object detection process. To put it simply, this python coding aims to read the text file and then create a voice narration.

B. Result

The system has been tested and its functionality has been demonstrated as per the design. The system has been able to operate as designed, thanks to the combination of software and hardware components. The system interface with Pi camera will capture the front environment and the data will be transferred to the processor to process the image. Object recognition methods will enable image processing, convert it to text, and use Google-Text-to-Speech to create the voice narration and send it to the user through an audio receiver.

The detected objects in frame are shown in Fig. 11.

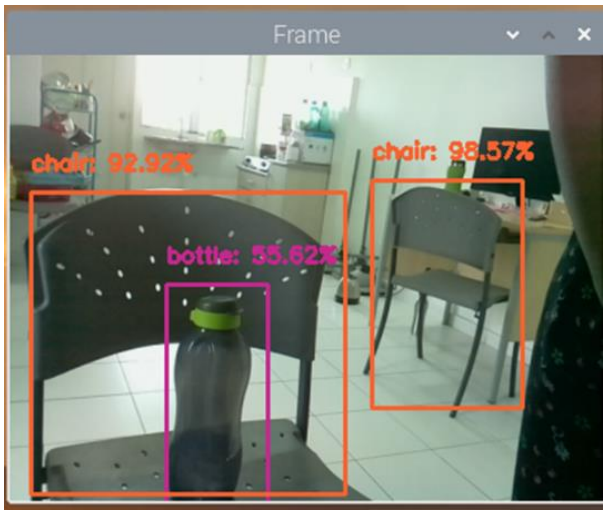


Fig. 11. Shows Objects Detected in Frame and its Confidence Percentage Value of the Detected Objects with its Label of Classes.

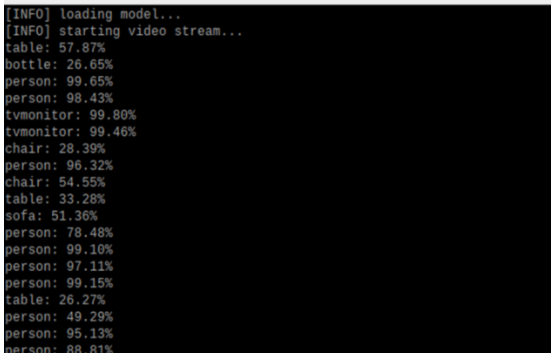


Fig. 12. Shows Label Classes will be Printed along with the Detection Confidence Percentage Value.

From Fig. 12, the results from the system are collected and the percentage of confidence value obtained, to show the objects detected along with its confidence percentage value. Confidence value is the probability that a bounding box containing an object and it is predicted by a classifier. The object in the bounding box would return many predictions, but out of those, most of them will have a very low confidence value associated. Hence, only predictions above 20% confidence is reported, as fixed in the python coding itself. That is how the object detection algorithm returns values after confidence thresholding, once the video stream starts in our system. As previously indicated in Fig. 11, the objects in the bounding box are correct, due to the quantifying the predictions.

To confirm the predictions, the correctness value of each object detection should be obtained. The measurement that determines the correctness of the bounding box is the Intersection over Union (IoU). IoU [21][22][23] is the ratio between the intersection, and the union of speculated boxes and ground truth boxes. The IoU's calculation is shown below in Fig. 13.

Consequently, the correct detections will be identified, and then its precision and recall will be calculated. To calculate precision and recall, the True Negatives, False Negatives, True Positives and False Positives will be identified. To obtain True Positives and False Positives, IoU will be used and the detection will be identified to determine whether it is correct (True Positive) or not (False Positive). The used threshold is 0.2, if IoU is > 0.2, it is considered a True Positive. Else, it will be considered as a False Positive. The COCO (Common Objects in Context) evaluation metric suggests measurements through various IoU thresholds.

To calculate the recall, the count of Negatives is required because not every part of the image in the video stream frame detected is an accepted object or is considered a negative. False Negatives will only be measured if the objects detected by our system are missed out. The recall is calculated as the ratio between the number of correct predictions (A) (True Positive) and the missed detections (False Negatives). The correct predictions for each class in the video stream will be commutated after calculation of IoU using the ground truth boxes for each positive detection box that the system has reported. So, with this, the IoU threshold (0.2). Therefore, the formulas will be as below.

$$Precision = \frac{TP}{(FP+TP)} \quad (1)$$

$$Recall = \frac{TP}{(TP/FN)} \quad (2)$$

Subsequently, the Mean Average Precision (mAP) is calculated in Table II. mAP is used in the domains; Information Retrieval and Object Detection. These two domains have separate ways to calculate mean average precision. Object detection of mAP is formalised in the PASCAL Visual Object Classes (VOC). PASCAL VOC provides a common dataset of images and annotations, as well as a standard evaluation to the vision and machine learning communities [24][25]. The average precision for all object types is shown in the table below. The PASCAL VOC dataset's mAP was found to be 0.665. The best mAP value at the moment is reported to be 0.739.

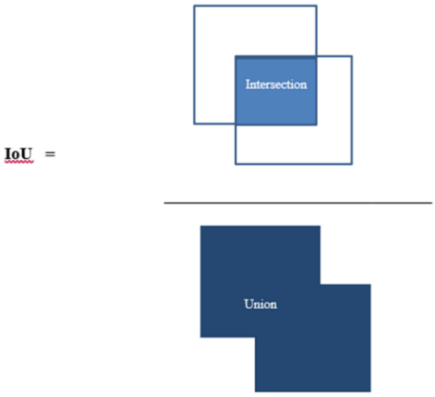


Fig. 13. IoU Calculation.

TABLE II. AVERAGE PRECISION FOR ALL CLASSES

No.	Class	Average Precision
1	Train	0.542
2	Bicycle	0.636
3	Dog	0.818
4	Diningtable	0.534
5	Aeroplane	0.727
6	Chair	0.909
7	Person	0.909
8	Tvmonitor	0.633
9	Bus	0.726
10	Sofa	0.710
11	Bird	0.727
12	Cow	0.632
13	Bottle	0.909
14	Pottedplant	0.359
15	Boat	0.544
16	Car	0.634
17	Cat	0.272
18	Sheep	0.633
19	Motorbike	0.724
20	Horse	0.726

VI. Conclusion

The system's goal of providing intelligent help for visually impaired people through real-time based object recognition has been successfully developed. Most of the important details in the general theory of design and execution have also been introduced throughout this article. From the theory to the practical realisation of this category of smart assistance for visually impaired people, these developments involve a variety of technical and coding details. From the testing and result analysis, the designed system's functionality is advanced and helps the visually impaired people to know what is in front of them. According to the data analysis based on Table II, the average precision for all classes is shown. Occasionally, detecting precision is not as precise as it should be, because the object is detected using values assigned by the system. Additional objects of comparable size or shape may also be detected with incorrect predictions. The strength of this system is users are able to listen to the voice narration audio that informs them what objects are in front of them. The Mean Average Precision (mAP) was calculated. The PASCAL VOC dataset's mAP was found to be 0.665. The best mAP value at the moment is reported to be 0.739.

The limitation of this system is it only has one pi camera interfaced to raspberry pi to capture video stream. So the scope only for blind people since the system included with pre-trained model that used for object detection. The most important recommendation for improvement, is about a future work development by implementing the Non-Maximum Suppression, making the regions more accurate. The object detection algorithm is good but not very accurate sometimes, because the regions reduce the ratio of algorithm. Furthermore, the development of this system should include a more pre-trained model in larger numbers. Lastly, the system can be improved by being cloud based, that way, all the data that had been captured will be saved in the cloud, and it will be easy for the user's guardian to acknowledge the details this system has generated, and can include localisation to know the location of the user travelled.

Acknowledgment

The authors would like to thank Centre for Research and Innovation Management (CRIM) for the support given to this research by Universiti Teknikal Malaysia Melaka (UTeM). We thank also those who contributed in any other forms for providing their continuous support throughout this work.

References

[1] M. Glatz, R. Riedl, W. Glatz, M. Schneider, A. Wedrich, M. Bolz, R. W. Strauss. "Blindness and visual impairment in Central Europe", PLoS one. 2022; 17(1):e0261897. https://doi.org/10.1371/journal.pone.0261897.

[2] SR. Flaxman, RRA. Bourne, S. Resnikoff, P. Ackland, T. Braithwaite, MV. Cicinelli, et al. Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. Lancet Glob Health. 2017;5(12):e1221–e34. pmid:29032195.

[3] TY. Wong, J. Sun, R. Kawasaki, P. Ruamviboonsuk, N. Gupta, VC. Lansingh, et al. Guidelines on Diabetic Eye Care: The International Council of Ophthalmology Recommendations for Screening, Follow-up, Referral, and Treatment Based on Resource Settings. Ophthalmology. 2018;125(10):1608–22. pmid:29776671.

[4] B. Thylefors, A. D. Negrel, R. Pararajasegaram, and K. Y. Dadzie. "Global data on blindness", Bulletin of the World Health Organization. vol. 73, no. 1, pp. 115–121, 1995. PMID: 7704921; PMCID: PMC2486591.

[5] A. Hydara, I. Mactaggart, S. J. Bell, J. A. Okoh, S. I. Olaniyan, M. Aleser, H. Bobat, A. Cassels-Brown, B. Kirkpatrick, M. J. Kim, I. McCormick, H. Faal, M. J. Burton. "Prevalence of blindness and distance vision impairment in the Gambia across three decades of eye health programming". The British Journal of Ophthalmology. 2021 Dec:bjophthalmol-2021-320008. DOI: 10.1136/bjophthalmol-2021-320008. PMID: 34949578.

[6] M. A. Ikbal, F. Rahman, M. R. Ali, M. H. Kabir and H. Furukawa, "Smart walking stick for blind people: An application of 3D printer", Proc. SPIE 10167 Nanosensors Biosensors Info-Tech Sensors 3D Syst., pp. 101670T, Apr. 2017.

[7] A. Krishnan, G. Deepakraj, N. Nishanth, and K. M. Anandkumar, "Autonomous walking stick for the blind using echolocation and image processing," 2016. 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), pp. 13-16.

[8] J. Bai, S. Lian, Z. Liu, K. Wang, and Di. Liu, "Smart guiding glasses for visually impaired people in indoor environment," IEEE Transactions on Consumer Electronics. vol. 63, no.3, pp. 258–266, 2017.

[9] Z. O. Abu-Faraj, E. Jabbour, P. Ibrahim, and A. Ghaoui, "Design and development of a prototype rehabilitative shoes and spectacles for the blind," 2012. 5th International Conference on Biomedical Engineering and Informatics, BMEI, pp. 795-799.

[10] S. Maidenbaum, S. Hanassy, S. Abboud, G. Buchs, D. R. Chebat, S. Levy-Tzedek, A. Amedi. "The "EyeCane", a new electronic travel aid for the blind: Technology, behavior & swift learning," Restorative neurology and neuroscience. vol. 32, no. 6, pp. 813–824, 2014.

[11] J. L. Robinson, V. Braimah Avery, R. Chun, G. Pusateri, W. M. Jay. "Usage of Accessibility Options for the iPhone and iPad in a Visually Impaired Population," Seminars in ophthalmology. vol. 32, no. 2, pp. 163–171, 2017.

[12] Grant, P.; Spencer, L.; Arnoldussen, A.; Hogle, R.; Nau, A.; Szlyk, J.; Nussdorf, J.; Fletcher, D. C.; Gordon, K.; & Seiple, W. "The functional performance of the BrainPort V100 device in persons who are profoundly blind," Journal of Visual Impairment & Blindness. vol. 110, no. 2, pp. 77–88.

[13] J. Spehr, "Object Recognition. In: On Hierarchical Models for Visual Recognition and Learning of Objects, Scenes, and Activities". Studies in Systems, Decision and Control, vol 11. Springer, Cham, 2015.

[14] B. Sandeep, S. Palaniappan, "Kinect language translator by using Google API," International Journal of Pharmacy and Technology. vol. 8, no. 4, pp. 20051-20060, 2016.

[15] J. Campos, S. Colteryahn and K. Gagneja, "IPv6 transmission over BLE Using Raspberry PI 3," 2018. International Conference on Computing, Networking and Communications (ICNC), pp. 200-204.

[16] M. Richardson and S. Wallace. "Getting Started with Raspberry Pi". O'Reilly Media, Inc., Sebastopol, 2012.

[17] V. Sati, S. M. Sánchez, N. Shoeibi, A. Arora, and J. M. Corchado, ''Face detection and recognition, face emotion recognition through NVIDIA Jetson nano,'' in Ambient Intelligence—Software and Applications, P. Novais, G. Vercelli, J. L. Larriba-Pey, F. Herrera, and P. Chamoso, Eds. Cham, Switzerland: Springer, 2021, pp. 177–185.

[18] W. A. Indra, A.F.M.F Ismail, Nurulhalim Hassim, M.H. Idris, S.G.Herawan, N.S. Zamzam, F. Zuska., "Feasibility of RF RSL for RF Energy Harvesting : A Case Study of Alor Gajah Area," 2021 IEEE 12th Control and System Graduate Research Colloquium (ICSGRC), 2021, pp. 24-28.

[19] K. D. Ismael and S. Irina, "Face recognition using Viola-Jones depending on Python," Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), vol. 20, no. 3, pp. 1513-1521, December 2020.

[20] N. Hassim, W. A. Indra, A.F.M.F Ismail, M.S.A.A Majid, S.G.Herawan, N.S. Zamzam, F. Zuska, "GSM900 Downlink Power Density Survey in Merlimau Area," 2021 IEEE 12th Control and System Graduate Research Colloquium (ICSGRC), 2021, pp. 99-103.

[21] F. Ahmed, D. Tarlow, and D. Batra, "Optimizing expected intersection-over-union with candidate-constrained CRFs," 2015. The IEEE International Conference on Computer Vision, pp.1850-1858.

[22] W. A. Indra, S. G. Herawan Industrial, N. S. Zamzam, S. b. Mohd Najib Fakulti, N. b. Hassim Fakulti and F. Zuska, "Development of A Guided Drone Powered by Radio Frequency Energy Harvesting," 2021 IEEE International Conference in Power Engineering Application (ICPEA), 2021, pp. 127-131.

[23] W. A. Indra, A. I. I. Jurjani, N. Hassim, S. G. Herawan, N. S. Zamzam and F. Zuska, "Digital TV Spectrum Survey for the Scope of Energy Scavenging in Jasin, Melaka," 2021 IEEE 17th International Colloquium on Signal Processing & Its Applications (CSPA), 2021, pp. 35-40.

[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” International Journal of Computer Vision. vol. 88, no. 2, pp. 303-338, 2010.

[25] A. -A. Nayan, J. Saha, K. Raqib Mahmud, A. Kalam Al Azad and M. Golam Kibria, "Detection of Objects from Noisy Images," 2020. 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), pp. 1-6.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 165 (2019) 259–269

Procedia Computer Science

www.elsevier.com/locate/procedia

INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, ICRTAC 2019

IoT based Assistive Device for Deaf, Dumb and Blind People

Karmel A*,Anushka Sharma, Muktak pandya, Diksha Garg

Vellore Institute of Technology, Chennai - 600 127, Tamil Nadu, India

Abstract

Focusing and addressing the problems faced by the differently abled people such as visually, audibly and vocally challenged, through a single device is a tough job. A lot of research has been done on each problem and solutions have been proposed separately. But not all of them are addressed together. The aim of the project is to create a single device solution in such a way that is simple, fast, accurate and cost-effective. The main purpose of the device is to make the differently abled people, feel independent confident by seeing, hearing and talking for them. The paper provides a Google API and Raspberry Pi based aid for the blind deaf and dumb people. The proposed device enables visually challenged people to read by taking an image. Further, Image to text conversion and speech synthesis is done, converting it into an audio format that reads out the extracted text translating documents, books and other available materials in daily life. For the audibly challenged, the input is in form of speech taken in by the microphone and recorded audio is then converted into text which is displayed in the form of a pop-up window for the user in the screen of the device. The vocally impaired are aided by taking the input by the user as text through the built-in customized on-screen keyboard where the text is identified, text into speech conversion is done and the speaker gives the speech output. This way the device speaks for the user.

Keywords: Raspberry Pi, Google APIs, Assistive Device, Text to speech, Deaf-Dumb-Blind people

1. Introduction

Approximately 1.3 billion people live with some sort of vision impairment out of which 188.5 million people have a mild vision impairment, 217 million have moderate to severe vision impairment, 36 million people are blind and the majority of people with vision impairment are over the age of 50 years. India is considered to home the largest number of blind people. Around 9.1 billion people are deaf and mute [1]. According to WHO, around 5% of the world's population or 466 million people suffer from disabling hearing loss. Technology is advancing day by day and during the last few decades, it has made our lives easier and convenient. But some- how the physically impaired part of our society has not been paid enough attention to. They are deprived of the advancements of science and still face plenty of problems in their day to day lives.

*Corresponding Author. Tel.: +91-9840507512
Email Address: karmel.a@vit.ac.in

10.1016/j.procs.2020.01.080

Communication is a major aspect of human lives. But there still exists a gap. Braille and sign language are the means of their communication but it is out of their comfort zone. They always have to learn these traditional modes of communication or they bank on support such as another person. This paper majorly focuses filling this gap by trying to make them feel independent and that they too can walk hand in hand with the other normal people. Raspberry Pi and Google API being the two pillars of this device, make it accurate, efficient and robust. The device consists of three major modules, each dedicated to the visually, audibly and vocally challenged. It uses Raspberry Pi supported by Google API as the main unit, and also consists of a camera, microphone, speaker and a screen. For the visually challenged, the inbuilt camera takes a picture of the writ- ten or printed document and this image is then converted into digital text using the Google Vision API. This text is then converted into audio using the TTS (Text to Speech) library and voice converted output according to the written document or book is obtained. The audibly challenged are aided by recording the speech or audio, converting it into text and displaying it on the screen for the user to read it. The device speaks for the vocally challenged as it provides the user with a customized keyboard on the screen where the user can type the message. This text is converted into speech using the TTS (Text to Speech) library and audio for the input given by the user is obtained in a synthesized voice.

2. Literature

Development of user centered interfaces and technologies have become crucial in the process designing for the differently abled people. Adding an extra element is just not enough to assist the use of technology for the visually disabled [5-7]. Many device-based hardware and software technologies exist to assist the visually disabled. They have functions like reading printed or writ- ten text, expanding characters on braille systems and machines Based on computer vision [3]. Prototypes that function with cell phone, cameras, help in processing images to identify patterns of movement, are applied for musicians who are blind [8,9].

AudioMUD [4] is a multiuser virtual environment exclusively made for the blind people and is associated with spoken cues. The original MUDs (Multi User Dimension) are generally text based and do not contain any sort of graphical interfaces. Users generally use MUD (Multi User Dimension) style games to per- form a set of actions in a virtual environment with a navigable space in the presence of direction, orientation and restrictions. There is high potential for the description of spaces and interactions due to its possible types of interaction and text based interface between players and virtual environment in AudioMUD with collaborative aspects. Their project focuses on the development of a server and client from scratch where the state of the world is stored in the server in such a way that when the server connects to the client, the state of virtual game is received and players can enter or exit anytime. The game starts when the blind user enters the IP name and server in the client, the player comes inside the kingdom of the human body like the respiratory system in a random location with attributes and can explore the system. L.Gonzalez et al [2] suggest a system for the visually disabled to enhance the quality of their life. The wearable system consists of facial recognition to recognize people's faces and can identify a person through prior system training using fisher faces algorithm, obstacles detection where the user wears the device which uses ultrasonic sensors to generate vibration signals that indicate an obstacle, email reader which accesses user's email using POP3 protocol and enables the user to listen to the email using headphones, medication reminder to remind the user about the medication prescribed, MP3 player as a source of entertainment enabling the user to listen to music. Anusha Bhargava et al [3] suggest a system using raspberry pi that uses image acquisition using interfacing a webcam, preprocessing of image to obtain the region of interest, template identification to detect characters and objects, converting image to text using OCR algorithm which scans image and gives a corresponding text output, and save the text data in a text file, and convert text to speech using E-speak for the blind user to hear the text.

Sign language which principally uses manual communication including hand movements, facial expressions to express, connects with people and convey their messages. Lorenzo Monti et al [10] have come up with a wearable device for the deaf-blind users called GlovePi to identify the person, number and position of people, and their facial expressions in front of the user. It mainly comprises of a gardener glove which is attached to capacitive touch sensor with raspberry Pi using a I2C interface. Using many to many architecture in order to include maximum amount of users into an account, the Glove enables the user to register on the server usually by sending a HTTP request and eventually the user is added on the server after which the server sends a updated list of all the connected users and thus uses peer to peer communication to send or receive messages. Amro Mukhtar Hussain et al [12] has designed a

mouth gesture recognition system using the help of an infrared sensor that collects the data from the audibly impaired person's mouth and detects the state of the mouth. They have designed three states: OSCS (Open Slow Close Slow), OSCF (Open Slow Close Fast) and OFCS (Open Fast Close Slow). When the sensor reaches its threshold, the sensor indicates and records the signal. Using different combinations, 27 patterns have been achieved which generated 26 alphabetic letters. The output of this proposed system depends on the light reflected from the object that the sensor subjected on, where the intensity supposedly gets affected by the surface color, shape and distance, after which the circuit gets the appropriate output analog voltage range.

Systems that suffice all solutions for the blind, deaf and dumb users in one compact device are rare to find. Kumar.K et al [1] have introduced an arrangement for the visually impaired can understand words using Tessaract which is an OCR (Optical Character Recognition) algorithm by python, vocally impaired can ex- press and communicate by text which is read through E-speak, and audibly impaired can hear by speech to text conversion using OpenCV. Rohit Rastogi et al [11] have put up an ideology that consists of a Sharon bridge which is a wearable technology that makes communication between differently abled on the extent of their capabilities. The Sharon Bridge comprises of small units to form a complete circuit to enable them to convey messages among the differently abled and their different combinations. It comprises of a sensor glove that is made up of arduino circuit board, tactile and flex sensors, and accelerometer which is used to convert the American sign language to audio that is further changed to text which is displayed on the LCD(Liquid Crystal Display) for the user, Arduino GSM(Global System for Mobile communications) shield to communicate over long distances using the internet and GPRS(General Packet Radio Service) wire- less network, Beagle bone that converts analog to digital and vice versa. It works in a way where the message to be sent is the input as text, audio or braille language which is converted to the respective forms for the disabled to hear, speak or see. For long distances, the input in converted and sent through wireless GSM network to the receiver but the user is supposed to possess a phone number. Sharon Bridge works for all combinations of the blind, deaf and dumb.

3. Design

The figure 1 shows the outline of the device. The raspberry Pi is the support system of the device which connects the camera, microphone, speaker and LCD display. The device works for the visually impaired as the camera clicks a picture of the document and the output is in audio format through the speaker, audibly impaired as the microphone takes the spoken words as input and displays it as text on the LCD display, and for the vocally impaired as the user types the message in the LCD and the speakers gives the output as an audio.

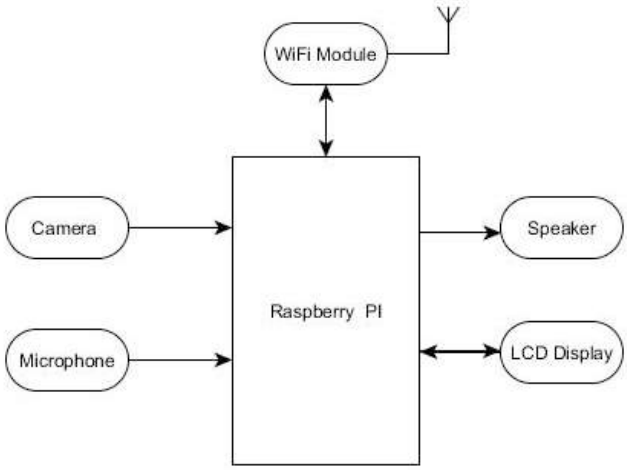


Figure 1 The system architecture with all main modules

3.1 Software Requirements

3.1.1 Google Cloud Vision API

The Google Cloud Vision API (Application Programming Interfacing) encapsulates powerful machine learning models in an easy to use REST API and enables developers and users to apprehend the content of an image. It is used for classification of images into thousands of categories, detecting individual objects and faces within images, and reading printed words contained within images. Optical Character Recognition (OCR) is used to enable the user to detect text within images, along with automatic language identification. Vision API supports a huge and broad set of languages. Initially Conventional neural network (CNN) based model is used to detect localized lines of text and generates a set of bounding boxes. Script identification is done by identifying script per bounding box and there is one script per box. Text recognition is the core part of the OCR which recognizes text from image as shown in Figure 2.

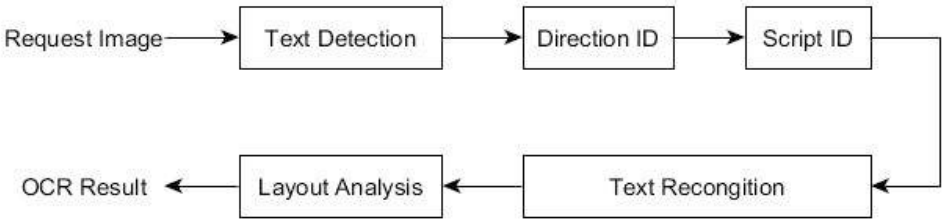


Figure 2 Architecture of Vision API converting Image to text

3.1.1 Tkinter

Various options for the development of graphical user interfaces are provided by python. Tkinter is the standard GUI (graphical user interface) provided as a library for python. GUI applications can be created in a faster and easier way using Tkinter, and it also provides a prevailing object-oriented interface to the Tk GUI toolkit.

3.1.2 Speech to Text:

Google cloud Speech to text aides the developers in the conversion of audio into text as it applies robust neural network models in a convenient API. It enables voice command and control and transcribes audio. It is capable of processing real-time streaming or pre-recorded audio using Google's ML technology. The accuracy is unparalleled as the most advanced deep learning neural network algorithms are applied by Google. It streams text results, returning text as it is recognized from audio stored in a file and is capable of long-form audio as shown in Figure 3.

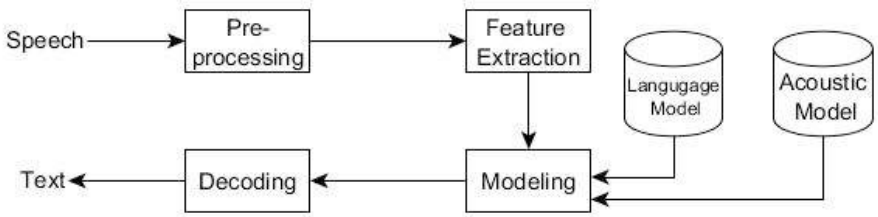


Figure 3 Google speech API converting speech to text

3.1.3 Text to speech

Google Text to Speech API is one of the several APIs available in python to convert text to speech as shown in Figure 4. It is commonly known as the gTTS API. It is an easy and efficient tool which converts entered text, into audio that can be saved as an mp3 file.

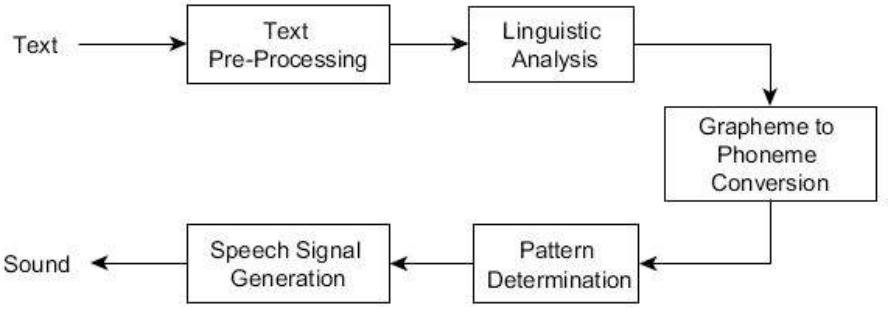


Figure 4 gTTS converting text to speech

3.1.4 Bitwise SSH

Bitwise SSH (Secure Shell) is one of the advanced and flexible SFTP protocol. The bitwise ssh helps us to securely connect with raspberry pi and access all the resources of raspberry pi. In addition, the user can transfer the files from localhost to raspberry pi; compile the programs; and provides a secure link for further connection.

3.2 Hardware Requirements

3.2.1 Raspberry Pi

Raspberry Pi, shown in Figure 5 is a low cost, credit card sized processor, which can easily perform all task we expect from a desktop. It is very easy to connect raspberry pi with computers and TVs. It also provides GPIO (General Purpose Input Output) pins to connect with other components. Because of this efficiency to intercommunicate with the cross-disciplinary domain, it has been used in a variety of projects. Raspberry pi operates in an open source environment such as Raspbian (Linux based operating system).

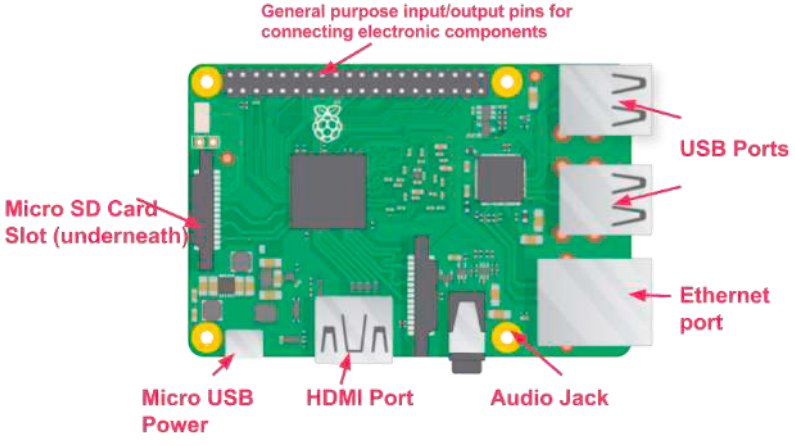


Figure 5 Raspberry Pi

Technical specification

- Broadcom Soc BCM2836 (CPU, GPU, DSP, SDRAM)
- 900 MHz quad-core ARM Cortex A7 CPU (ARMv7 instruction set)
- Broadcom VideoCore IV @ 250 MHz GPU
- 1 GB MEMORY (shared with GPU)
- 4 USB ports
- 17 GPIO(General Purpose Input Output) Peripherals plus specific functions, and HAT ID bus
- 15-pin MIPI camera interface (CSI) Video input connector
- HDMI video outputs, composite video (PAL and NTSC) via 3.5 mm jack
- I²S audio input
- Analog audio output via 3.5 mm jack; digital via HDMI and I²S
- MicroSD for storage
- 10/100Mbps Ethernet speed
- 800 mA power rating (4.0 W)
- 5 V power source via MicroUSB or GPIO(General Purpose Input Output) header
- 85.60mm × 56.5mm
- Weighs 45g (1.6 oz)

3.2.2 Camera Module

The camera used by project is a C310HD Logitech webcam, shown in Figure 6 with a resolution of 720p/30fps. The images taken are crisp and contrasted. This camera fits perfectly in the project as it adjusts to the lighting conditions to produce brighter contrasted images. It uses a universal clip to attach itself firmly to the device. It is small, adjustable and agile and is therefore handy in the project.

Technical specifications

- Max Resolution: 720p/30fps
- Lens technology: standard
- Focus type: fixed focus
- Field of View: 60°
- Built-in mic: mono
- Cable length:1.5 m
- Universal clip fits laptops, LCD or monitors

Figure 6 Logitech C310HD Webcam

3.2.3 Screen display

The project consists of a 5inch resistive touch with a high hard- ware resolution and HDMI Interface specially designed for the Raspberry Pi, shown in Figure 7. It has a resistive touch control It is compatible and has a direct connects with any revision of the existing Raspberry Pi. It provides drivers and the backlight can be turned

on or off for the lower power consumption. According to the requirements of the project, a keyboard has been hardwired in this 5-inch dis- play for the vocally challenged to type their text in the screen.

Figure 7 Waveshare 5' inch display

Technical Specification

- Drivers provided (works with your own Raspbian/Ubuntu/Kali/Retropie)
- HDMI interface for displaying, no I/Os required (however, the touch panel still needs I/Os)
- High-quality immersion gold surface plating

3.2.4 Microphone

A mini portable high quality USB microphone, shown in Figure 8 is used in the project. It is a noise cancelling microphone which filters out un- wanted background noise. It comes as a brownie point for the project as it is portable, compact and easy to use. It can be made more efficient according to the user or background by increasing the gain control or capture for better accuracy.

Technical Specification:

- 4.5V Working voltage
- Weight 99.8 g
- 2cm x 2cm x 0.5cm in size

Figure 8 USB Microphone

4. Implementation

The device has been created by formulating a unique design for assisting the differently abled people. It has been divided into three modules for enhancing the experience of the user with the device. The device consists of three modes and a three-way slider to change mode. Each mode is separately dedicated for the blind, deaf and dumb respectively in the device. The device is designed to make the user feel individualistic, self-reliant and self-sufficient. The gist of design of the device is in the figure 1. The main component of the device is the raspberry Pi.

4.4.1 Blind Module

The figure 9 represents the methodology of this module which consists of three steps.

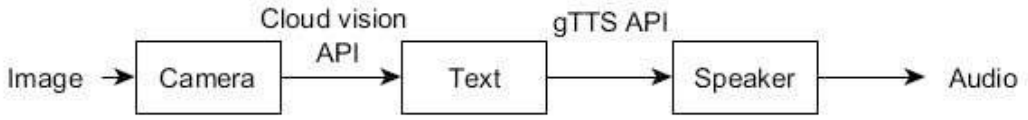


Figure 9 Working of Blind Module

Step 1.For the module to work, the three-way slider is set to the blind mode. The camera connected to the raspberry Pi of the device takes a picture of the written document or book placed on the holder of the device.

Step 2.The picture is saved in JPEG format and is passed to the Google Cloud Vision API to be converted to text where the API extracts the text to be converted.

Step 3.The extracted text then gets converted into speech using the gTTS API and the required text is thus converted to the audio format.

Step 4.This audio is given as an output by the high quality speaker connected to the Raspberry Pi and thus the device enables the visually impaired person to understand the written document or book through the audio.

4.4.2 Deaf module

The audibly impaired can virtually hear using this device as it enables them to read, what is being spoken. The figure 10 describes the respective procedure.

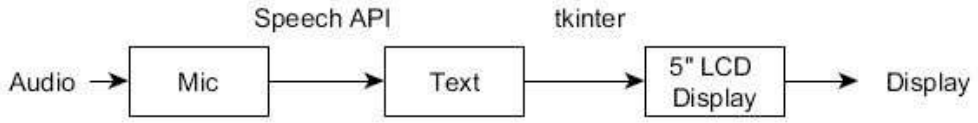


Figure 10 Working of Deaf Module

Step 1.The three-way slider is set to the deaf mode. The audio or the words being spoken to the user, who in this case might be a deaf person, are recorded as input by the USB Microphone connected to the Raspberry Pi of the device and is saved as a file in mp3 format.

Step *2*.This audio file is passed to the Google Speech API which converts the audio into text for the user to understand.

Step 3.The converted text is then displayed on the 5 inch HDMI LCD screen available in the device, as a pop up window exclusively created using python tkinter for this module. This way the user understands everything that is being spoken to him quickly and efficiently. To change modes, the slider can be set accordingly.

4.4.3 Dumb module

This module makes the device handy for the vocally disabled as it enables them to vocalise words by typing it on the screen. The figure 11 explains the methodology.

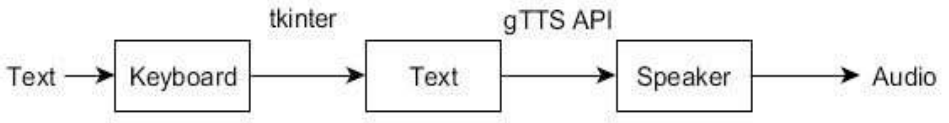


Figure 11 Working of Dumb Module

Step 1.When the three-way slider is used to set the device on the dumb mode, a pop up is displayed along with a customized keyboard which has been created using python tkinter, in it on the HDMI screen connected to the Raspberry Pi.

Step 2.The user who possibly is vocally impaired can type whatever he wants to convey using the keyboard in the screen as text

Step 3.The typed text is converted into audio format using the gTTS API and the audio file of the required text is obtained.

Step 4.The high quality speaker connected to the Raspberry Pi in the device plays this audio file thus vocalising the message given by the impaired person.

Step 5.Modes can be changed in the device according to the convenience of the user.

5. Results

The figure 12 shows the working of the blind module as it reads the text taken as a picture in figure 13 and visually impaired will hear this text through the speakers.

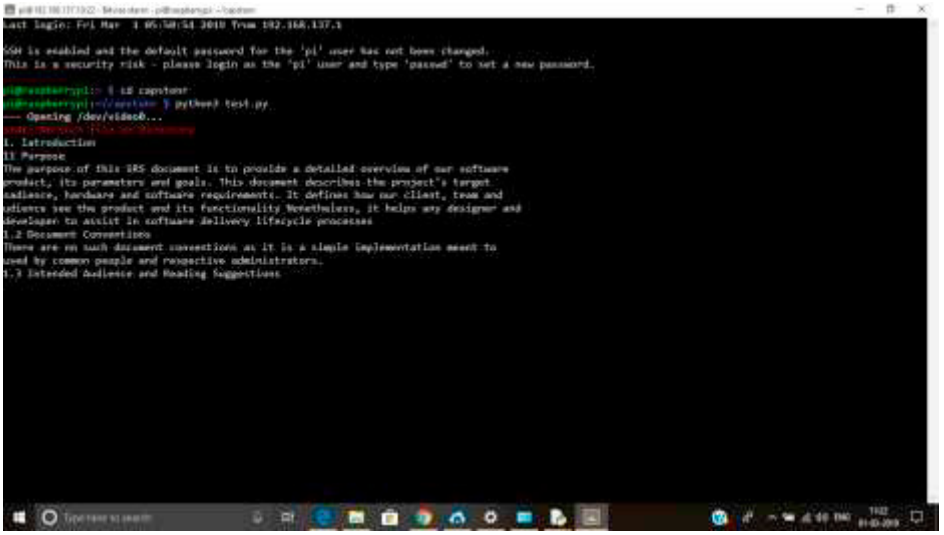


Figure 12. Conversion of image to text using Google Vision API

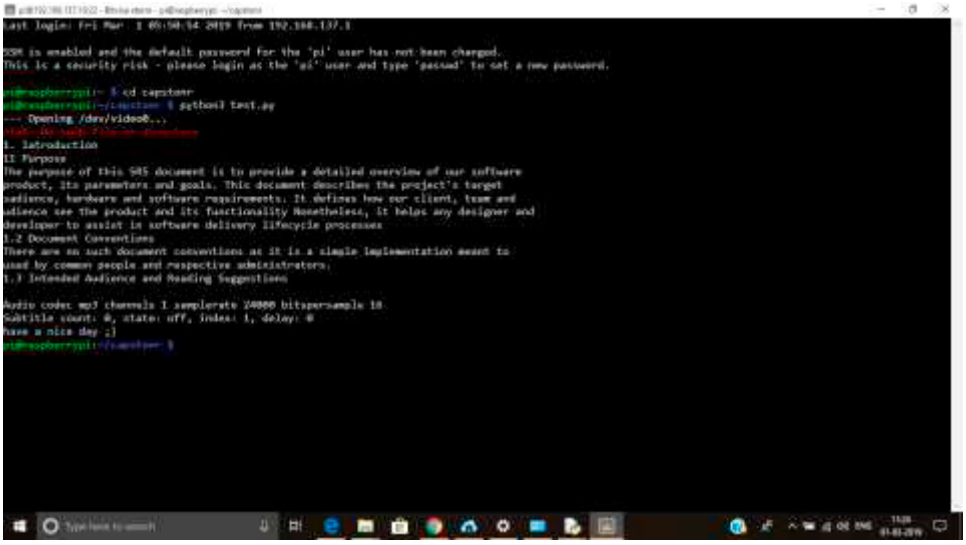


Figure 13. Conversion of gained text from image to Audio

Figure 14 shows how the audibly impaired can read as the audio or spoken words “Muktak doing testing for module 2” are identified as and converted to text.

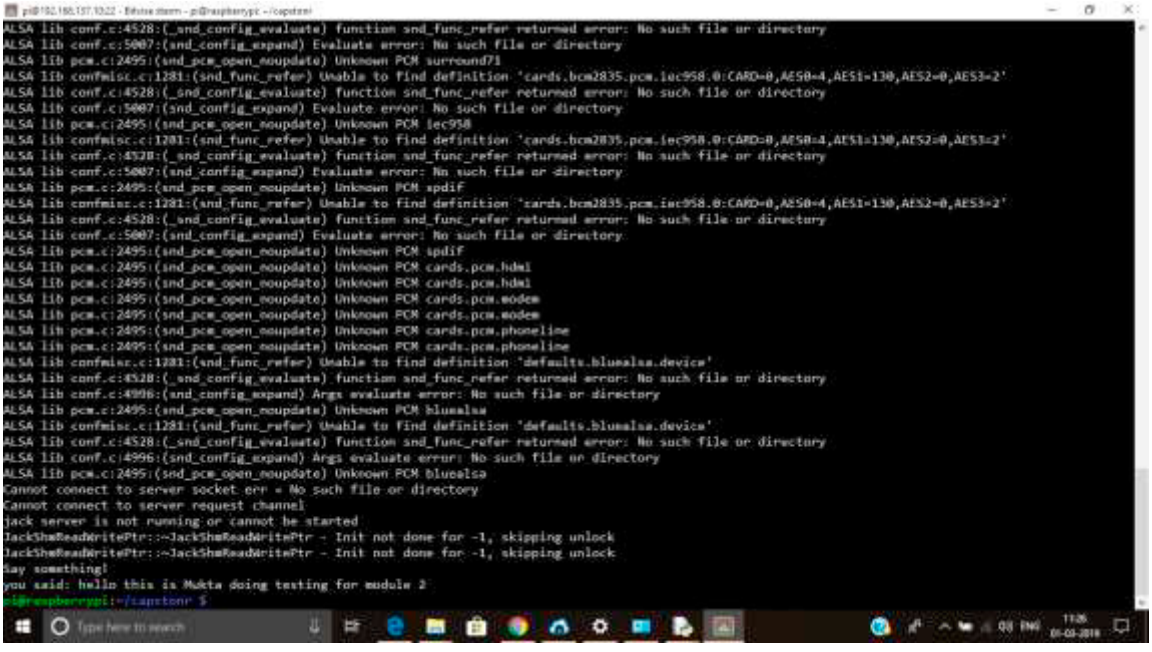


Figure 14. Conversion of Audio to text

Vocally impaired type a message in the keyboard of the screen as shown is figure 15 and this text gets converted to speech which the speaker gives as an output.

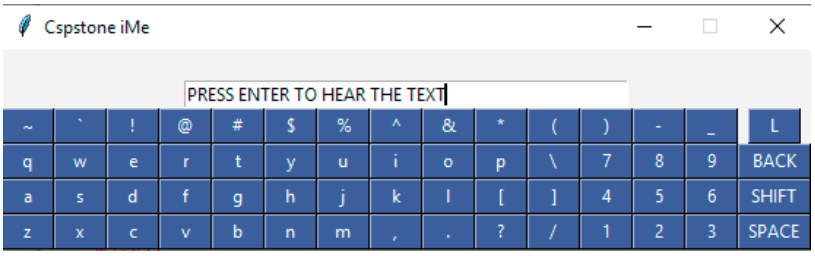


Figure 15 On-screen Keyboard (text to speech)

6. Conclusion

Through this paper, an unprecedented prototype has been created to aid the visually, vocally and audibly disabled. This project not just focuses on empowering and facilitating the differently abled, it is also compact and resource saver. The overall cost has been cut down by eliminating braille books and the energy spent in understanding them. It is a less costly solution, as all the components used in the device are cost effective and efficient. The latest and most trending technology makes this device portable, adaptable and convenient. The device proposed in this paper can be a major help in solving a few of the many challenges faced by the differently abled. To further extend the project, the device can be made more compact and wearable to make it easy for the user to use.

References

[1] N. K., S. P. and S. K., Assistive Device for Blind, Deaf and Dumb People using Raspberry-pi, Imperial Journal of Interdisciplinary Research (IJIR), 3(6), 2017 [Online]. Available: https://www.onlinejournal.in/IJIRV3I6/048.pdf.

[2] L. González-Delgado, L. Serpa-Andrade, K. Calle-Urgiléz, A. Guzhñay-Lucero, V. Robles-Bykbaev and M. Mena-Salcedo, "A low-cost wearable support system for visually disabled people," 2016 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), Ixtapa, 2016, pp. 1-5. doi: 10.1109/ROPEC.2016.7830606

[3] Anusha Bhargava, Karthik V. Nath, Pritish Sachdeva & Monil Samel (2015), International Journal of Current Engineering and Technology, E-ISSN 2277– 4106, P-ISSN 2347– 5161

[4] J. Sanchez and T. Hassler, "AudioMUD: A Multiuser Virtual Environment for Blind People," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, 15(1), pp. 16-22, March 2007. doi: 10.1109/TNSRE.2007.891404

[5] M. Lumbreras and J. Sánchez, “Interactive 3-D sound hyperstories for blind children,” in Proc. ACM-CHI ’99, Pittsburgh, PA, 1999, pp. 318–325.

[6] R. McCrindle and D. Symons, “Audio space invaders,” in Proc. ICDVRAT 2000, Alghero, Sadinia, Italy, Sep. 23–25, 2000, pp. 59–65.

[7] T. Westin, “Game accessibility case study: Terraformers-Real-time 3-D graphic game,” in Proc. ICDVRAT 2004, Oxford, UK, 2004, pp. 120–128.

[8] Y. H. Lee and G. Medioni, “Rgb-d camera based wearable navigation system for the visually impaired,” Computer Vision and Image Understanding, vol. 149, pp. 3–20, 2016

[9] J. Bajo, M. A. Sanchez, V. Alonso, R. Berj ´ on, J. A. Fraile, and J. M. ´ Corchado, “A distributed architecture for facilitating the integration of blind musicians in symphonic orchestras,” Expert Systems with Applications, 37(12), pp. 8508–8515, 2010.

[10] L. Monti and G. Delnevo, "On improving GlovePi: Towards a many-to-many communication among deaf-blind users," 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, 2018, pp. 1-5.doi: 10.1109/CCNC.2018.8319236

[11] R. Rastogi, S. Mittal and S. Agarwal, "A novel approach for communication among Blind, Deaf and Dumb people," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015, pp. 605-610.

[12] A. M. Hassan, A. H. Bushra, O. A. Hamed and L. M. Ahmed, "Designing a verbal deaf talker system using mouth gestures," 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), Khartoum, 2018, pp. 1-4. doi: 10.1109/ICCCEEE.2018.8515838

Optical Braille Recognition Based on Semantic Segmentation Network with Auxiliary Learning Strategy

Renqiang Li[1] Hong Liu[1*] Xiangdong Wang[1] Jianxing Xu[2] Yueliang Qian[1]
[1] Beijing Key Laboratory of Mobile Computing and Pervasive Device,
Institute of Computing Technology, Chinese Academy of Sciences
[2]Northwestern Polytechnical University
[1]{lirenqiang,hliu,xdwang,ylqian}@ict.ac.cn
[2]xjx_david@outlook.com

Abstract

Optical Braille Recognition methods usually use many designed steps, such as image de-skewing, Braille dots detection, Braille cell grids construction and Braille character recognition, which are less robust for complex Braille scenes. This paper proposes an optimal semantic segmentation framework BraUNet to directly detect and recognize Braille characters in the whole original Braille images. BraUNet adds extra auxiliary learning strategy to UNet network, which uses long-range connections of feature maps between encoder and decoder to get more low-level features. And auxiliary learning strategy can combine multi-class Braille characters segmentation with Braille foreground extraction, which can improve the feature learning ability and the Braille segmentation performance. Then morphological post-processing is used on semantic segmentation results to get the final individual Braille character regions. Experimental results show the proposed framework is robust, effective and fast for Braille characters segmentation and recognition on both complex double sided Braille image dataset and handwritten Braille image dataset.

1. Introduction

According to the latest WHO survey [1], there are about 1.3 billion people with some degree of vision impairment in the world. Braille is a basic writing language for the visual impaired to learn knowledge and communicate with each other. Braille documents are constructed by Braille characters or Braille cells, which are lied in Braille cell rows and Braille cell columns according to detailed Braille arrangement rules. Each Braille cell is made up of six raised or flat Braille dots arranged in three rows and two columns. So there are 64 different Braille character classes including the empty Braille cell.

Recently, many Optical Braille Recognition (OBR) systems are proposed, which focus on detecting Braille cells from Braille document images and converting them into corresponding natural language characters [2]. OBR systems are useful and meaningful to protect and republish early precious Braille books, recognize handwriting Braille documents and automatically evaluate examination papers in the special education fields, which are now mainly processed manually.

Braille cell detection and recognition is the basic technology in OBR systems. Existing methods are mainly based on traditional image processing techniques [3, 4] and machine learning methods [5, 6]. These methods are usually based on several complex steps [2], such as image de-skewing, Braille dots detection, Braille cell grid construction, and Braille character recognition. These steps are designed carefully to achieve satisfactory performance according to Braille appearance and arrangement rules.

While in some complex situations, such as the irregular arrangement of Braille cells suffered from Braille printing noise, image acquisition noise, and complex handwritten Braille images. It's difficult to design appropriate rules to convert the Braille dots to Braille characters in above complex situations. Especially for handwritten Braille images, Braille rows or paragraphs may have different skew angles. These different skew angles make it difficult to perform image de-skewing using traditional arrangement rule based methods.

With the great success of deep learning on ImageNet in 2012 [7], deep learning has made impressive progress on many difficult tasks such as image classification, object detection and semantic segmentation. However, there are very few applications in optical Braille recognition. Some existing methods mainly focus on classification of the cropped Braille character images using CNN networks. These methods cannot directly process the whole Braille image, which is difficult to apply in real applications.

This paper focuses on general and effective Braille characters detection and recognition on the whole original Braille images. We directly consider Braille characters or Braille cells as targets to detect and segment instead of Braille dots. For this task, we propose a robust OBR framework based on semantic segmentation network BraUNet and morphological post-processing method. The framework can process the original Braille images by end-to-end way without traditional complex steps.

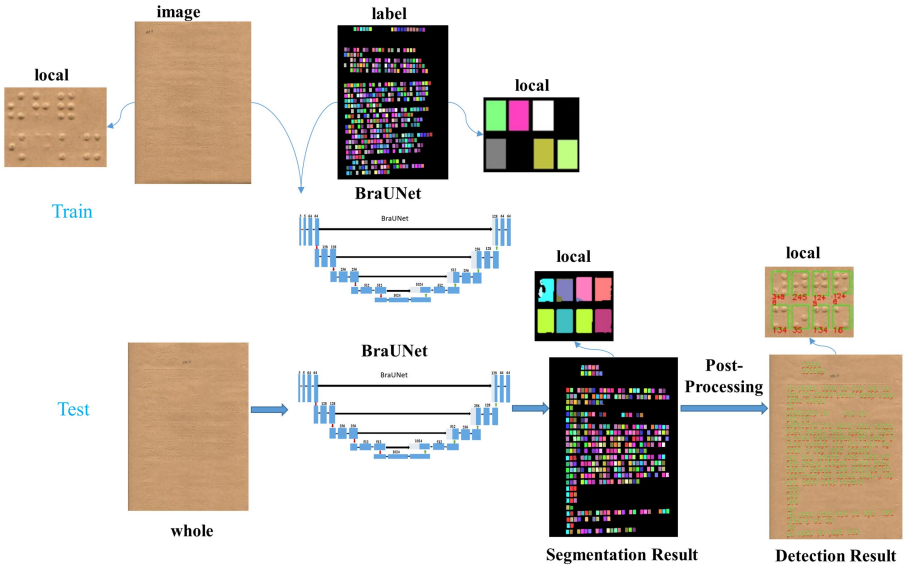


Figure 1: Our proposed framework for Braille characters detection and recognition.

The proposed BraUNet combines multi-class Braille characters segmentation task and auxiliary Braille cell foreground segmentation task to supervise the network learning process. And U-Net structure [8] uses long-range connections of feature maps between encoder and decoder to get more low-level features. These strategies can improve the feature learning ability and the whole Braille character segmentation performance. We further use morphological post-processing on semantic segmentation results to get the final Braille character regions.

The experimental results show the proposed framework is more general, robust and effective for Braille characters recognition on the public double-sided Braille image dataset and collected handwritten Braille document images.

2. Related work

Optical Braille Recognition systems are developed from 1990s [9], which can be grouped into three categories.

Traditional image processing techniques are widely used [2, 3, 4]. Antonacopoulos et al introduced a local adaptive thresholding method to segment the Braille image into three parts including shadows, light and background, and then identify Braille recto dots and verso dots by the combination rules of these parts [3]. They also constructed a Braille grid based on Braille arrangement rules to convert the Braille dots to Braille characters. However, these methods are sensitive to segmentation thresholds and designed rules.

In order to overcome the above shortcomings, some methods used machine learning techniques to recognize Braille dots. Li et al [6] adopted the cascaded classifier with Haar to quickly detect the Braille dots. Li and Yan [5] used SVM with a sliding window strategy to recognize Braille dots. Recently, Li et al [6] proposed a two-stage learning framework for double-sided Braille images recognition. A cascaded classifier with Haar is used to quickly detect the Braille dots in the first stage. Then the detected Braille dots are used for images de-skewing and constructing Braille cell girds. They also used multiple SVM classifiers with HOG or LBP features to further classify each intersection on the grid in the second stage. Namba and Zhang [11] used cellular neural network to only classify 10 Braille numbers on the cropped Braille character images.

Recently, some researchers use deep learning methods to classify the cropped Braille characters. Li et al [12] used stacked denoising autoencoder to classify 10 Braille numbers. Kawabe et al [13] trained a CNN network AlexNet to classify Braille recto dots, verso dots and background.

Above image segmentation based methods and traditional machine learning based methods for OBR may contain several steps, such as image preprocessing, de-skewing, Braille dots detection, Braille grid construction, and Braille characters recognition. These methods are limited by multi-stage processing and designed rules, which are difficult to apply to complex Braille images. Some neural network and deep learning based methods only classify cropped Braille characters with few classes or Braille dots. So far, there is a lack of research on the recognition of 64 braille characters on the whole Braille image, especially for complex double-sided Braille images and handwritten Braille images.

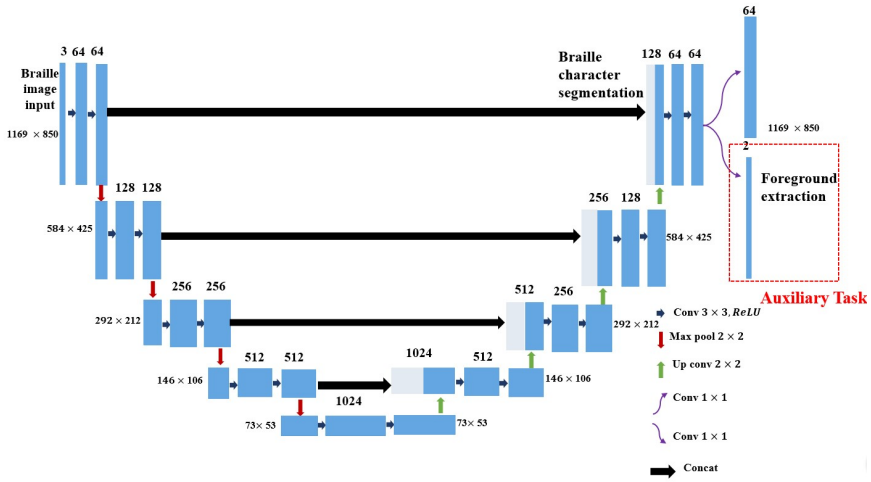


Figure 2: The architecture of our network BraUNet.

3. Our proposed work

This paper applies the semantic segmentation method in natural image analysis into the task of Optical Braille Recognition. A general and robust OBR framework is proposed based on the semantic segmentation network BraUNet and morphological post-processing techniques. Compared with the existing methods, the proposed framework can directly detect and recognize Braille characters in the whole original Braille images without relying on several steps and complicated rules.

Fig. 1 shows our proposed framework. Original Braille images and the corresponding pixel-level annotations of Braille characters with 64 classes are input into the semantic segmentation network BraUNet for training and testing. Then morphological post-processing techniques are used to refine segmentation results and get the final Braille character bounding box regions. We will introduce our framework in detail in following sections.

3.1. Network architecture of BraUNet

We propose an optimal network BraUNet for our OBR task, which is based on U-Net [8] and adopts extra foreground extraction based auxiliary task to improve the feature learning ability. U-Net structure uses long-range connections between the encoder and decoder to connect the feature maps with corresponding size, which can enhance the accuracy of low-level edge prediction and get the refined pixel labels with size of original image. U-Net is widely used in medical image segmentation due to its simple structure and good performance.

Fig. 2 shows the architecture of our network BraUNet. The left part is a contracting path used to extract high-level semantics features and the right part is an expansive path used to recover the original size gradually. The skip connection to recover the details of the segmentation result from the higher level is also adopted in our network.

This paper inputs the whole Braille images and their labeled images with Braille characters to train semantic segmentation network as Fig. 1 shows. While pixel-level segmentation for 64 Braille characters is not robust in some complex scenes due to acquisition noise, background disturbance and irregular Braille arrangement. Some segmentation errors may influence the shape of Braille characters and make wrong recognition for Braille characters. If we regard each Braille character or Braille cell as the single foreground object, the task to extract foreground from background may be simpler than multi-class semantic segmentation.

So inspired by the idea of [14], we add foreground extraction of the Braille cell as an auxiliary task at the end of the framework in our task. We train the multi-class Braille characters segmentation and the auxiliary Braille cell foreground segmentation task in the same network and calculate the corresponding loss. As shown in Fig.2, we simultaneously output the results of the foreground extraction and Braille character segmentation at the end of the network. The result of foreground extraction is used to improve the feature learning ability and supervise accurate generation of Braille cell boundary during training.

3.2. Annotations for Braille Images

The semantic segmentation network requires pixel-level annotations for model training. To alleviate the labelling workload, we simply use the bounding boxes to annotate the Braille characters on the original whole Braille images and directly convert the bounding boxes into the pixel-level

annotation results. We create an empty annotation image L(i, j) with the same height and weight of the original image. For each of the annotated Braille characters, we can get a bounding box R_k and its class type c. For all position i, j in the R_k, we assign each pixel p(i, j) in image L(i, j) with the value c. Here c is an integer from 0 to 63 representing 64 types of Braille character. And the background and empty Braille characters are all assigned as 0. In this way, we can easily get pixel level annotations for Braille characters in Braille images. For auxiliary learning task, we just remain the background pixel as 0 value and change the rest pixel with 1 value as Braille cell foreground.

3.3. Network training

Different from the existing methods using the high-resolution Braille image such as 200dpi [6, 10] or even 600dpi [5], we only use 100dpi Braille images as the network input, which can greatly reduce the data storage, transmission and inference time. In the data preprocessing stage, we firstly down sampling the input 200dpi RGB Braille color image in DSBI dataset from 2338×1700 to 1169×850 pixels, which is the original size of the image scanned by 100 dpi.

For 64 classes Braille characters semantic segmentation task, we use the combined Dice loss [15] and Cross Entropy (CE) loss as our loss function. They are defined as follows:

$$L_{Dice} = \frac{1}{C-1}\sum_{c=1}^{C-1}\left(1 - \frac{2\sum_i^N p^c(i)g^c(i) + \gamma}{\sum_i^N p^c(i) + \sum_i^N g^c(i) + \gamma}\right) \quad (1)$$

$$L_{CE} = -\frac{1}{N \times C}\sum_{c=0}^{C-1}\sum_{i}^{N} g^c(i) log p^c(i) \quad (2)$$

$$L = L_{Dice} + \alpha L_{CE} \quad (3)$$

Where $p^c(i)$ and $g^c(i)$ denote the predicted value and the ground truth respectively at the position i of the whole image, C denotes the class number, which is 64 in our paper, and N denotes the total number of pixels of the whole Braille image. For smoothing purposes, we add a γ factor to both the nominator and the denominator to avoid denominator is 0. In our experiments, α is set to 1.

For the auxiliary foreground extraction task, we also use the combined Dice loss and CE loss, except that C is set to 2. Finally, the total loss is defined as follows:

$$L_{total} = L_{multi\_class} + \beta L_{auxiliary} \quad (4)$$

The weight β is set to 1 in our experiments. We train the network for 70 epochs in our experiments with the optimizer Adam and learning rate 1e-4. And the best model is selected according to the Dice value on the validation set during training process.

3.4. Post processing

In the test stage, we input the original whole Braille image into our trained BraUNet and get the initial pixel-level semantic segmentation result for Braille characters. Due to the segmentation noise and the small spacing between adjacent Braille characters, some segmented Braille characters areas are connected each other. We further use morphological processing methods to get the final bounding box of each Braille character. We binarize each class type of Braille character based on segmentation results. For each binary image, an erosion operation is used to reduce the adhesion between adjacent Braille characters, and the connected component analysis is used to extract the contour of each Braille character. To reduce noise, some small areas are removed. Finally, we take the bounding box of each contour as the detected and recognized results of Braille characters.

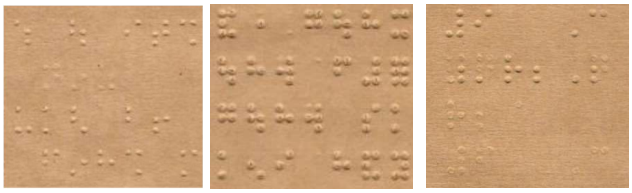
4. Experiments and analysis

4.1. Dataset

Double Sided Braille Image dataset DSBI. We adopt public Braille image dataset DSBI [10] to evaluate our method, which contains 114 double-sided Braille images from several Braille books. Li et al [6] proposed a two stage learning framework TS-OBR for Braille character recognition based on Braille dots detection on DSBI with 200dpi resolution. They used 26 Braille images for training, which is sufficient for detecting and classifying Braille dots. But for training deep learning models, such as the semantic segmentation model with 64 classes of Braille characters, 26 images are not enough for high performance. In our paper, we divide DSBI into three subsets including training, validation and test set with 74, 10 and 30 images respectively. Images from each Braille book in DSBI are proportionally sampled to construct the three subsets. And recto Braille characters and verso Braille characters detection and recognition are all evaluated in our experiments. We use 200dpi resolution of Braille image for TS-OBR and 100dpi for U-Net and BraUNet methods which are obtained by down sampling images directly.

Braille Answer Sheet Dataset BAS. To evaluate our method for handwritten Braille recognition, we collect some Braille answer sheets from a special education school. These answer sheets are all written by different students using some certain Braille boards. It is challenge for handwritten Braille recognition.

The biggest problem in handwritten Braille is that Braille characters in different lines or paragraphs in the same page may have different skew angles. This is usually caused by the location change of the Braille writing board, which makes rule-based methods fail. On the other hand, different students may have different writing habits which leads to various appearance of Braille dots as shown in



(a) Tiny Braille dots (b) Thick dots (c) Erased dots

Figure 3: Some handwritten Braille dots and Braille characters.

Fig.3(a) and Fig.3(b). Another difficulty is that many students modify certain Braille characters by directly erasing one or more Braille dots in one Braille character, which is more difficult to distinguish whether these Braille dots are erased or not by visual information as Fig.3(c). We collect total 50 Braille answer sheets images from 32 students. These handwriting Braille images are single sided with recto Braille dots.

4.2. Metrics

Dice value [8] is adopted to evaluate the performance of semantic segmentation results with or without auxiliary task of U-Net model. For Braille characters detection and recognition performance, we also use the Precision, Recall and F1 values in [10] as the evaluation metrics. In addition, we use the Intersection over Union (IOU) to evaluate the degree of overlap of two Braille character boxes. For each predicted Braille character box, we use all the ground truth of Braille character boxes to calculate the IOU value. If the maximum IOU value is greater than threshold T, we assume that the predicted Braille character box is correct. In our experiments, T is set to 0.5. We denote the number of Braille characters correctly classified as TP, the number of Braille characters misclassified but with the correct position as FP, the number of Braille characters with the wrong position as WP, and the number of Braille characters missed as TN. The Precision, Recall and F1 values can be defined as follows:

$$\mathrm{Pre} = \frac{TP}{TP + FP + WP} \tag{5}$$

$$\mathrm{Rec} = \frac{TP}{TP + FN} \tag{6}$$

$$\mathrm{F1} = \frac{2 \times Pre \times Rec}{Pre + Rec} \tag{7}$$

4.3. Experimental results

4.3.1. Experiment setting

To evaluate our proposed method, we conduct two experiments to compare semantic segmentation network BraUNet with U-Net, and existing method TS-OBR [6].

U-Net means the original network [8] and BraUNet means our U-Net with auxiliary foreground segmentation task. U-Net and BraUNet are all deep learning based methods. We also compare our method with the recent method TS-OBR [6] based on traditional machine learning method. TS-OBR uses a cascaded classifier to quickly detect the Braille dots and then de-skews image to construct a Braille cells grid, and further adopts SVMs to classify each intersection on the grid for Braille dots. This method relies on the construction of Braille cells grid, which is sensitive to noise and irregular arrangement. We retrain and test TS-OBR method on our newly divided sets of DSBI with 200dpi resolution of Braille images.

All the models of U-Net, BraUNet and TS-OBR are trained only on the public double-sided Braille dataset DSBI. We evaluate the semantic segmentation and detection performance including recto Braille characters and verso Braille characters on DSBI.

We further use single-sided Braille answer sheet dataset BAS as test set to evaluate the generalization ability of Braille character detection and segmentation methods, which are trained on DSBI dataset.

4.3.2. Results of semantic segmentation

Table 1 shows the results of pixel-level semantic segmentation performance of Braille characters for U-Net and BraUNet on both DSBI and BAS datasets. Fig.4 shows the Dice columns figure on DSBI and BAS datasets. Dice value is used to evaluate the result of semantic segmentation performance without any post-processing. On DSBI dataset, the Dice value of our BraUNet improves 5.48% for recto Braille characters than U-Net, and improves 3.17% for verso Braille characters. From table 1, we can also find the Dice value of BraUNet drop from 0.9508 on DSBI to 0.9048 on BAS, which maybe some data distribution of single-sided handwritten Braille characters are different from those of double-sided printed Braille documents. This problem can be resolved by adding training sample from BAS for fine tuning model. And on BAS dataset, the Dice value of BraUNet improves 5.43% compared with U-Net for recto Braille characters segmentation. Fig.4 also shows above conclusion.

Table:1 Semantic segmentation results of Braille characters for U-Net and BraUNet.

Dataset	Type	Method	Dice
DSBI	Recto Braille character	U-Net	0.8960
		BraUNet	**0.9508**
	Verso Braille character	U-Net	0.9040
		BraUNet	**0.9357**
BAS	Recto Braille character	U-Net	0.8505
		BraUNet	**0.9048**

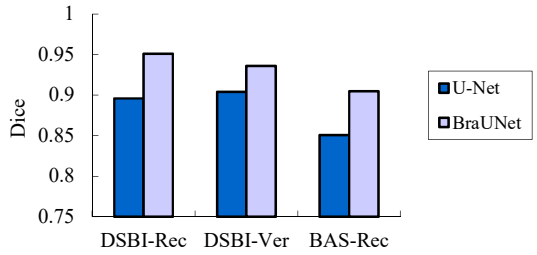
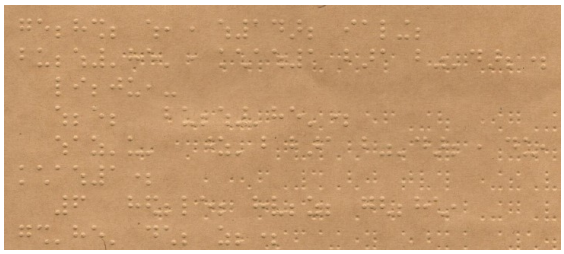
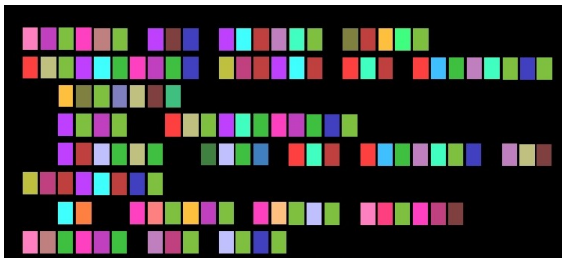


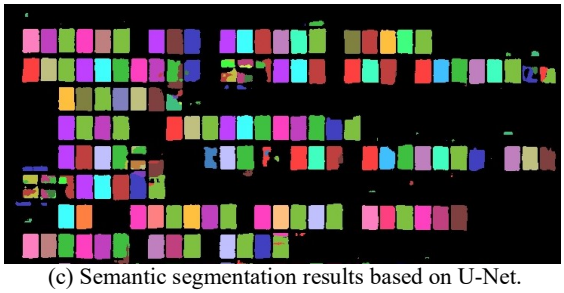
Figure 4: Braille character segmentation performance.



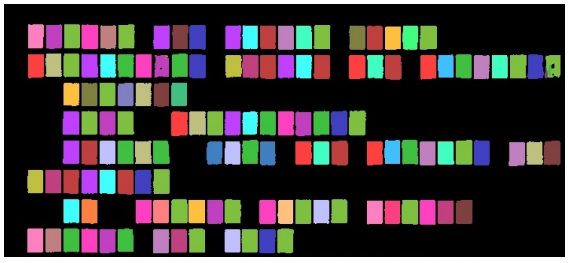
(a) Local region of source Braille image in DSBI.



(b) Manual labels with 64 classes of Braille characters.



(c) Semantic segmentation results based on U-Net.



(d) Semantic segmentation results based on BraUNet.

Figure 5: Semantic segmentation results from a local regions of Braille images in DSBI. Different color means different class of Braille characters.

The above results show the effectiveness and generation performance of our optimal for both recto and verso Braille character segmentation.

Fig.5 shows some semantic segmentation results of recto Braille characters from a local part of one Braille image on DSBI dataset. Fig.5(a) is a local region of an original double-sided Braille image and Fig.5(b) is the ground truth label of recto Braille character with 64 different colors. Fig.5(c) is the semantic segmentation results of U-Net model, which shows many noise and wrong segmentation pixels especially on the edge region of each Braille character. Fig.5(d) is the results of BraUNet, which adds auxiliary foreground segmentation task to U-Net. It's clear that the outline of Braille characters are improved compared with the results of U-Net in Fig.5(c). And the segmentation and recognition noises and errors of some Braille characters in Fig.5(c) are improved in Fig.5(d).

4.3.3. Results of Braille characters detection

Table 2 shows the detection performance of Braille characters for U-Net, BraUNet and TS-OBR on both DSBI and BAS datasets. The metrics of precision, recall and F1 are used to evaluate detection performance. Based on above semantic segmentation results, we further use morphological post-processing methods in section 3.4 to get the final bounding box of each Braille character.

On DSBI dataset, despite the disturbance of the Braille dots on the back page and the insufficient amount of training data, original U-Net network still can achieve 0.9751 and 0.9768 F1 values for the recto and verso Braille character recognition. These results show that the end-to-end semantic segmentation model is useful for Optical Braille Recognition. With the help of auxiliary task, our BraUNet network finally gets 0.9966 and 0.9893 F1 values for recto and verso Braille characters respectively.

On BAS dataset, we get 0.9425 and 0.9638 F1 value for recto Braille characters detection for U-Net and BraUNet respectively. Nearly 2% improvements of F1 value shows our auxiliary task is effective, which improves the feature learning ability and segmentation performance for complex double-sided Braille and handwritten Braille images.

Table 2 also shows the comparative results of our deep learning method and existing TS-OBR method [6]. The F1 values are similar for BraUNet and TS-OBR on DSBI dataset including recto and verso Braille character detection. While TS-OBR used 200dpi resolution of Braille images which are double size of width and height compared with our 100dpi for BraUNet.

On BAS dataset, TS-OBR just gets the 0.9408 F1 value, which is about 2.3% lower than BraUNet. This maybe there are many irregular arrangements of Braille characters in handwritten Braille documents, which will bring error to construct the accurate Braille cell grid using TS-OBR, so some Braille characters could not be obtained correctly.

Table 2: Comparative results of Braille characters detection.

Dataset	Type	Method	Dpi	Pre-%	Rec-%	F1
DSBI	Recto Braille character	U-Net	100	98.35	96.69	0.9751
		BraUNet	100	99.43	99.88	**0.9966**
		TS-OBR	200	99.28	99.96	0.9962
	Verso Braille character	U-Net	100	98.31	97.06	0.9768
		BraUNet	100	98.81	99.05	0.9893
		TS-OBR	200	98.44	99.70	**0.9906**
BAS	Recto Braille character	U-Net	100	92.90	95.64	0.9425
		BraUNet	100	93.50	99.44	**0.9638**
		TS-OBR	200	89.47	99.19	0.9408

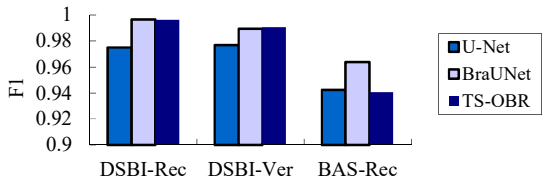


Figure 6: Braille character detection performance.

While our proposed BraUNet is more robust and can end-to-end get the Braille character recognition results without multiple complex steps and designed rules in TS-OBR. Fig.6 shows the F1 columns figure of U-Net, BraUNet and TS-OBR for Braille characters detection on DSBI and BAS datasets, which also shows the effectiveness of our optimal BraUNet compared with U-Net and TS-OBR methods.

We implement our BraUNet framework using the deep learning framework PyTorch with one GPU 1080Ti. The average processing time of Braille character segmentation and recognition is about 0.25s for one Braille image.

5. Conclusion

This paper introduces an effective Braille segmentation and recognition framework for whole Braille images. We propose an optimal semantic segmentation network BraUNet with auxiliary learning task by end-to-end way. This auxiliary learning task can combine multi-class Braille characters segmentation and Braille cell foreground extraction, which can improve the feature learning ability and segmentation performance of Braille characters. The morphological processing algorithms are used to get the final Braille detection results. The proposed framework is effective and general, which can directly detect and recognize Braille characters in the original Braille images without relying on Braille dots detection, image de-skewing and Braille arrangement rules. The experimental results on public double-sided Braille image dataset and collected Braille answer sheet dataset show the robustness and effectiveness of BraUNet. In the future, we will collect more complex Braille images to test and improve OBR performance.

Acknowledgement

This work is supported in part by Beijing Haidian Original Innovation Joint Foundation (L182054).

References

[1] Visual impairment and blindness, World Health Organization, 2018.

[2] Samer Isayed and Radwan Tahboub. A review of optical braille recognition. In WSWAN, pages 1–6. IEEE, 2015.

[3] Apostolos Antonacopoulos and David Bridson A robust braille recognition system. InDAS, pages 533–545, 2004.

[4] Majid Yoosefi Babadi and Shahram Jafari. Novel grid-based optical braille conversion: from scanning to wording. International Journal of Electronics,98(12):1659–1671,2011.

[5] Lie Li and Xiaoguang Yan. Optical braille character recognition with support-vector machine classifier. In ICCASM2010, volume 12, pages V12–219. IEEE, 2010.

[6] Renqiang Li, Hong Liu, Xiangdong Wang, and Yueliang Qian.Effective optical braille recognition based on two-stage learning for double-sided braille image. In PRICAI, 2019.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural net-works. In NIPS, pages 1106–1114, 2012.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, pages 234–241, 2015.

[9] Jan Mennens, Luc Van Tichelen, Guido Francois, and Jan J.Engelen. Optical recognition of braille writing. In ICDAR, pages 428–431, 1993.

[10] Renqiang Li, Hong Liu, Xiangdong Wang, and Yueliang Qian. DSBI: Double-sided braille image dataset and algorithm evaluation for braille dots detection. In ICVIP, pages 65–69. ACM, 2018.

[11] Michihiro Namba and Zhong Zhang. Cellular neural network for associative memory and its application to braille image recognition. In IJCNN, pages 2409–2414, 2006.

[12] Ting Li, Xiaoqin Zeng, and Shoujing Xu. A deep learning method for braille recognition. International Conference on Computational Intelligence and Communication Networks, pages 1092–1095. IEEE, 2014.

[13] Hiroyuki Kawabe, Yuko Shimomura, Hidetaka Nambo, andShuichi Seto. Application of deep learning to classification of braille dot for restoration of old braille books. In ICMSEM, pages 913–926. Springer, 2018.

[14] Wenjia Wang, Junxuan Chen, Jie Zhao, Ying Chi, XuansongXie, Li Zhang, and Xian-Sheng Hua. Automated segmentation of pulmonary lobes using coordination guided deep neural networks. ISBI 2019, pages 1353–1357, 2019.

[15] Carole H Sudre, Wenqi Li, Tom Vercauteren, et al. Generalized dice overlap as a deep learning loss function for highly unbalanced segmentations. In Deep learning in medical image analysis and multimodal learning for clinical decision support, pages 240–248, 2017.

International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 2 Issue 2, February 2015

ISSN: 2394-4404

Object Recognition In Mobile Phone Application For Visually Impaired Users

Kharmale Arati

Jori Sayali

Dangare Sushanta

Ahire Harshita

Danny Pereira

Comp Dept, Govt College Of Engineering And Research (Awasari),
Pune, India

Abstract: Blind people face a number of challenges when interacting with their environments because so much information is encoded visually .There are many problems when blind people need to access visualizations such as images, objects, information in the form of text etc. Many tool and technologies seek to help blind people solve these problems by enabling them to query for information such as color or text shown on object. Blind use Braille technique to read. Also there are many applications like screen reader which help them to read. But there is a need of special training to use these techniques and also they are not so much portable. In this work we describe main features of software modules developed for Android smart phones that are dedicated for the blind users. The main module can recognize and match scanned objects to a database of objects, e.g. food or medicine containers. The two other modules are capable of detecting major colors and locate direction of the maximum brightness regions in the captured scenes. We conclude the paper with a short summary of the tests of the software aiding activities of daily living of a blind user.

Keywords: accessible environment, blindness, image recognition

I. INTRODUCTION

The blind and the visually impaired face diverse kinds of life challenges that normally sighted people take for granted. As far as out-door activities are concerned the blind indicate difficulties in safe and independent mobility depriving them of normal professional and social life. Then the issues dealing with communication and access to information are pointed out. Here a significant help is offered by software applications for computers and touch-screen devices equipped with speech synthesizers that enable browsing the internet and access to text documents. Finally, the common problem experienced by the blind are the so called activities of daily living. Activities in this area and the corresponding aids can be subdivided into the following main components:

- ✓ personal care, i.e. labeling systems, health care monitoring and the use of medicines,
- ✓ timekeeping, alarm and alerting, e.g. tools for controlling household appliances, smoke aond monoxide detectors,
- ✓ food preparation and consumption; utensils of special construction and talking (or sonified) kitchenware (e.g. liquid level indicators),
- ✓ indoor environmental control and use of appliances, i.e. accessible systems providing feedback information about in a format suitable for a blind user,
- ✓ money and finance; high- and low-tech solutions making shopping and money operations possible for the blind.

In this work we focus our attention on the group of daily living tasks related to personal care systems, in particular, systems that enable the visually impaired to identify objects, e.g. food packs, medicine containers and other items. Currently available solutions can be grouped into the two major groups:

- ✓ low-tech labeling systems in which labels are attached to objects, e.g. with tactile signs or text messages in,
- ✓ high-tech systems, that employ 1-D and 2-D barcodes, talking labels or radio-frequency identification devices (RFID).

Both systems, however, require attaching special tags or visual signs to the objects. Consequently, they can be costly, since such systems need to be regularly maintained to keep them up to date. In this communication we report a solution aimed at aiding the visually impaired in color detection, light direction detection and recognition of objects. The system is based on a dedicated image recognition application running on an Android system smart phone. Image recognition results are communicated to the blind user by means of pre-recorded verbal messages.

II. INDENTATIONS AND EQUATIONS MATHEMATICAL MODEL

A. MAPPING

1]

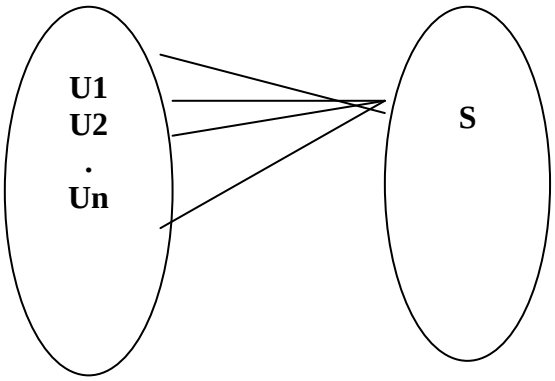


Figure 1: one system will be used by many blind users

2]

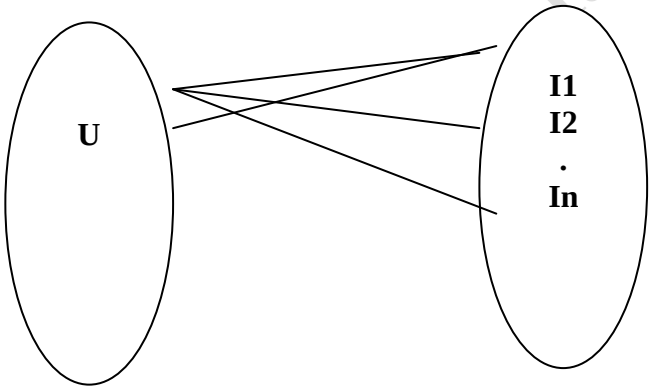


Figure 2: one user will submit many object images to the system

B. SET THEORY

Our system can be represented as a set
System S = {I,O,C}
Where,
I=set of inputs
O=set of outputs
C = set of constraints
Input
Input I= {Object Image}
Object Image = {Object Image1, Object Image2,…, Object Imagen }
Output
Output O = {Recorded verbal message}
Constraint
C1 = "The system should have an huge set of object images with its respective verbal message stored in database".

III. MODULES USED

In the developed software tool for Android smart phones we propose the three following image processing modules:

- ✓ Color detection module
- ✓ Object recognition module
- ✓ Light source detector.

An advantage of the algorithm we proposed for object recognition model is its scale and orientation invariance.

A. THE COLOR DETECTION MODULE

The color detection algorithm works on images taken with an automatic flash with the smallest resolution possible (320x480 pixels). The RGB color images are converted into the HSI (Hue Saturation Intensity) color images. This color space enables to represent the color in a single parameter, i.e. the H component, whereas the S component is the saturation parameter of the recognized color. We tested two different approaches. In the first approach, in the HSI color space, the average value of the color (the H component) is determined for the photo taken. The average color is compared with a predefined reference table of colors and the color of the photographed object is determined. If the image is too dark or too bright (this decision is made based upon the saturation and intensity components) a special warning message is generated. In the second approach a special color histogram is computed. Each image pixel is represented in the HSI space and allocated to a predefined color histogram bin. If there is a significant disproportion between the first and the second most frequent image color the most frequent one is communicated to the user only. If the occurrence frequency of more than one color is similar, the application informs the user about mixture of colors, e.g. by voicing the message "yellow-red color".

B. THE LIGHT DETECTOR MODULE

The idea of light detector module was proposed by one of the blind users. The application operates in real time and is based on the content of the camera's preview image. In the first step, the average brightness of the centre part of the image is calculated. Next, an audio signal with a frequency dependent on the average brightness is generated. The brighter the image the higher is the frequency of the generated sound. The application proved suitable in robust localization of light sources, for example streetlights or a lamp in the room.

C. IMAGE RECOGNITION ALGORITHM

The goal was to design an application which would allow to recognize objects from images recorded by the camera of a mobile device. The object recognition algorithm should be insensitive to image registration parameters, i.e. scale, rotation and lighting conditions. Moreover, the recognized object should be robustly detected and localized in the image context (e.g. among other similar objects). The SIFT (Scale-Invariant Feature Transform) proposed in was applied in the developed application. The SIFT is considered as a very power computer vision algorithm for detecting and describing local image features. SIFT allows to compute feature descriptors strongly independent on the image registration conditions. These descriptors are further used to recognize objects in the proposed application.

Figure 3: Object recognition with the help of SIFT

IV. SYSTEM ARCHITECTURE

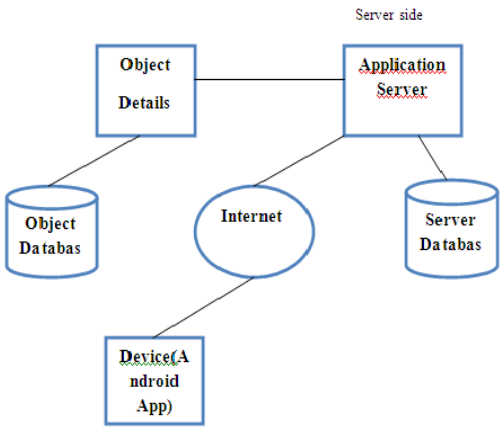


Figure 4: System Architecture

V. CONCLUSION

Image processing is a very vast area and it is among rapidly growing technologies today, with its application in various aspects of a research. The text present on an image gives the important information about an image. In this project we extract that text from the image and convert it into speech. The project is developed mainly for the blind people.
The converted voice will help them to identify the object whose image is captured by them directly as input to software and can listen it. In this way, we can help to blind people in some sort of their activities such as identifying the objects, reading text etc. In addition to this, it will also help to people having reading disabilities in reading text. We can use this software on android mobile as well as on the computer system having web-cam

VI. ACKNOWLEDGEMENT

The satisfaction that accompanies that the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. We are grateful to our project guide and HOD, Department of Computer Engineering Mr. Pareira D. J. for the guidance, inspiration and con-structive suggestion that helpful us in the preparation of this project. Gracious gratitude to all the faculty of the department of Computer Engineering valuable advice and encouragement.

REFERENCES

[1] Strumillo P. (2012) Electronic navigation systems for the blind and the visually impaired, Lodz University of Technology Publishing House (in Polish).
[2] Hersh M., Johnson M. (Eds.) (2008) Assistive technology for visually impaired and blind people, Springer, London
[3] Gill J. (2008), Assistive devices for people with visual impairments.In: Helal S., Mokhtari M. and Abdulrazak B. (ed) The engineering handbook of smart technology for aging, disability, and independence, J Wiley and Sons, Inc, Hoboken, New Jersey, pp. 163-190.
[4] Onishi J., Ono T. (2011) Contour pattern recognition through auditory labels of Freeman chain codes for people with visual impairments, Inter-national IEEE Systems, Man, and Cybernetics Conference, Anchorage, Alaska USA, pp. 1088-1093.
[5] Introducing Mobile Speak. http://www.codefactory.es/en/products.asp?id=316. Accessed 25th February 2013
[6] LookTel Recognizer. http://www.looktel.com/recognizer. Accessed 25th February 2013.
[7] Matusiak K., Skulimowski P. (2012), Comparison of Key Point Detectors in SIFT Implementation for Mobile Devices, ICCVG Lecture Notes in Computer Science vol. 7594, Springer, pp. 509-516.

ISSN: 2394-4404

Optical Braille Recognition Using Object Detection Neural Network

Ilya G. Ovodov
ELVEES RnD center, JSC
Zelenograd, Russia
iovodov@elvees.com

Abstract

Optical Braille recognition methods generally rely heavily on a Braille text's geometric structure. They run into problems if this structure is distorted. Thus, they find it difficult to cope with images of book pages taken with a smartphone.

We propose an optical Braille recognition method that uses an object detection convolutional neural network to detect whole Braille characters at once. The proposed algorithm is robust to deformations and perspective distortions of a Braille page displayed on an image. The algorithm is suitable for recognizing braille texts captured with a smartphone camera in domestic conditions. It can handle curved pages and images with perspective distortion. The proposed algorithm shows high performance and accuracy compared to existing methods.

Additionally, we produced a new dataset containing 240 photos of Braille texts with annotation for each Braille letter. Both the proposed algorithm and the dataset are available at GitHub.

1. Introduction

The embossed Braille alphabet was invented in 1824 and served as the primary way of writing and reading for blind people for many years. Recent advances in technology provide blind people with many new opportunities for receiving and transmitting information. Still, reading printed Braille and writing in Braille continues to be an important communication method for them. Moreover, Braille is often used for communication between blind and sighted people. In particular, the situation is quite common when sighted teachers work with blind students, and they have to deal with textbooks and student works written in Braille.

Braille is an alphabet made of dots embossed on paper for tactile recognition. Still, sighted people who deal with Braille texts usually do not have enough tactile skills to read the text with their fingers and have to read it with their eyes. For sighted people, Braille text looks like a lot of tiny white bulging points on a white background, so its visual recognition is very tedious. Reading double-side printed Braille is especially difficult. Text printed on a back of a page looks like dented dots. Tactilely these dots are almost invisible and do not interfere with the sensation of front side convex points. Still, visually they are hardly distinguishable from them (see Figure 1). So, reading such a text with the eyes is especially difficult. The use of technical Braille recognition tools, particularly optical recognition methods, can greatly facilitate this work. Thus, optical Braille recognition (OBR) methods have been developing since at least the 1980s [2].

Figure 1. Example of a page Taken with double-sided printing.

Each letter or other character is represented in a Braille text with several (1 to 6) bulging points located in a 2x3 grid. So, 63 different characters can be encoded. The distance between adjacent symbols is slightly larger than between two columns of points in one symbol. The character width, the spacing between characters, the spacing between lines, and the characters' places on a line are constant for each Braille document. Thus, the Braille characters' dots are located at nodes of a fixed grid (Figure 2). The use of this geometric structure is critical to most existing methods of Braille recognition. However, this significantly limits their applicability. These algorithms require either a scanner or special photographing conditions to ensure accurate alignment of a Braille page.

This work aims to provide the recognition method applicable to Braille text images obtained on a mobile phone camera in the domestic environment. The grid Braille char-

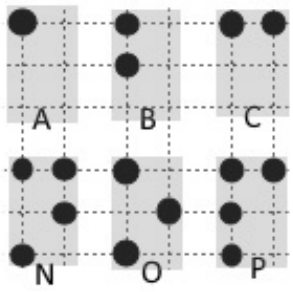


Figure 2. Schematic arrangement of Braille text.

acters are attached to can be significantly distorted due to: a) perspective distortions caused by the fact that the sheet is not perpendicular to the camera optical axis, b) the paper sheet curvature on open book spread images. Also, different areas of the sheet can have significantly different lighting (Figure 3).

To cope with these problems, we proposed an optical Braille recognition algorithm based on object detection convolutional neural networks. The proposed algorithm is robust to distortions of a Braille page image described above. So, the algorithm is suitable for recognizing braille texts captured with a smartphone camera in everyday conditions. The proposed algorithm has shown high performance and accuracy compared to existing methods.

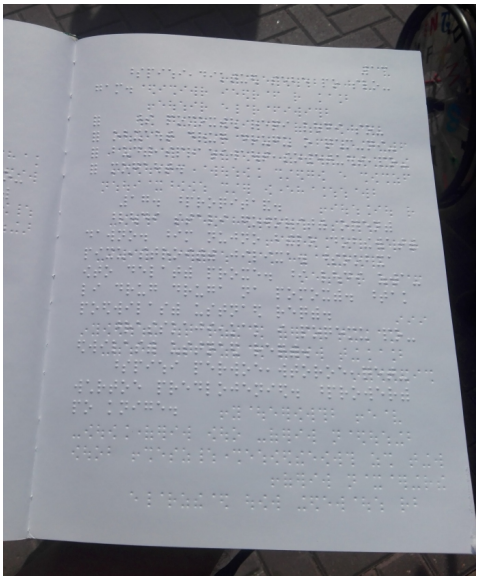
Additionally, we faced the lack of publicly available datasets that can be used to train a deep neural network and evaluate recognition algorithms' accuracy. The only available dataset is limited and does not cover the cases described above. We produced a new dataset containing 240 photos of Braille texts with annotation for each Braille letter.

2. Related work

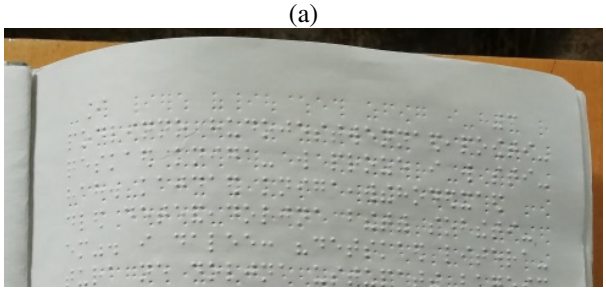
2.1. Optical braille recognition approaches

The main approach to optical braille recognition consists of sequentially performing Braille points search, characters grid restoration, including compensation for possible image rotation, points grouping into characters, and, finally, character decoding. The survey papers [21, 5] consider all algorithms following this sequence of steps.

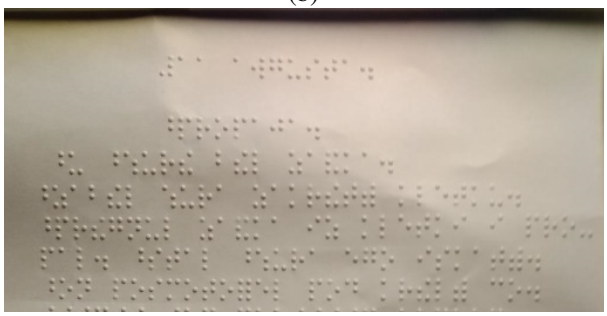
The simplest approach to point detection is thresholding. Zhang and Yoshino [24] use a dynamic local threshold. To detect dots on double-sided braille and distinguish between the front side and reverse side dots, Antonacopoulos and Bridson [1] use detection of light and dark areas of dots and segment image into areas of three classes: bright, dark, and background. They use static thresholds relative to the average brightness level in the vicinity of the point. Morgavi and Morando [16] use a simple neural network to



(a)



(b)



(c)

Figure 3. Examples of Braille text smartphone photos with perspective distortion (a), page curvature (b), and deformation (c).

find points. Venugopal-Wairagade [22] fulfills circle detection using Hough transform. Perera, Wanniarachchi [17] use HOG and SVM, R.Li *et al*. [10] - Haar detector and Adaboost., R.Li *et al*. [11] - Haar detector and Adaboost for primary dot detection and HOG, LBP, and SVM for final dot detection after grid restoration.

The next step is to restore the grid to which the points are anchored and possibly compensate for the sheet rotation. For this purpose can be used linear regression [17], Hough Transform [1, 6], or coordinates distribution density dur-

ing step-by-step rotation of the image can be investigated to find optimum rotation angle [11]. Sometimes, after the image de-skewing, points are searched again using information about their possible position [1, 11].

Although some methods assume the grid deformation, implying a change in the pitch between different lines ([11] and others), it is assumed that the grid lines are straight on the entire sheet and parallel.

Since the mentioned works essentially rely on snapping Braille points to the grid, they mostly work with images obtained with a scanner. In this case, the necessary image correction is reduced to de-skewing, which makes the grid lines vertical and horizontal. Only a few works declare the purpose of OBR on images from a smartphone ([24, 22]), but the methods described by them are still based on the presence of a rectangular grid on a sheet.

Although convolutional neural networks (CNN) made tremendous advances in image recognition in recent years, the use of deep learning and neural networks for OBR is sparse. There are only a few papers on the use of fully connected neural networks. Morgavi, Morando [16], and Ting Li *et al*. [13] use a simple neural network to find points, Subur *et al*. [20] - to find the symbol value using the points found by image segmentation. Kawabe *et al*. [6] use it for separating front and back points when recognizing two-sided braille. Only R.Li *et al*. [12] presented at CVPRW 2020 work based on segmentation neural network with modified UNet architecture. They used a neural network to determine areas occupied by Braille characters and recognize these characters. To determine the positions of individual characters based on segmentation results, subsequent post-processing is required.

2.2. Optical braille recognition methods accuracy evaluation

While various works provide different quantitative accuracy characteristics of proposed algorithms, sometimes quite high, comparing algorithms with each other encounters at least two obstacles:

- Until recently, there was no open dataset to compare different algorithms. Quality values provided in papers were measured on proprietary, not published datasets, so it is mostly impossible to compare published results of different works.

- Often works that use the general pipeline described above (i.e., points detection-grid restoration-grouping points into symbols and decoding), quality indicators are given only for the point recognition stage. It prevents comparing the performance of these algorithms with algorithms that do not have a separate point recognition stage, such as our algorithm.

The only publicly available DSBI dataset with braille text was published by Li *et al*. [10]. It contains 114 pages of scanned two-sided braille texts, divided into the train (26 pages) and test (88 pages) sets. All pages are carefully aligned during scanning. A grid of points for the front and back sides has been calculated, rotation required to bring the grid to a vertical-horizontal orientation has been calculated. The annotation is made by specifying the rotation angle, the coordinates of vertical and horizontal grid lines after rotation, and a list of braille characters referenced to this grid's nodes. All texts are in Chinese, but the Braille alphabet has the same structure in all languages, and this dataset can be used regardless of language.

Although this dataset does not look large enough and variable enough to provide full training of recognition algorithms (only 26 pages in the training set), it allows you to compare different approaches to the problem. The authors of [10] compared the accuracy of algorithms based on image segmentation (Antonacopoulos *et al*., [1]), Haar features and Adaboost (Viola & Jones [23]), and their algorithm (Li *et al*. [11]). However, they provide only the accuracy of point detection. Only in [12] they provided accuracy metrics of their algorithm, estimated not at the dot level, but at the character level.

2.3. Object detection convolutional neural nets

The use of convolutional neural networks (CNN) in computer vision and, in particular, in object detection has made tremendous strides in recent years. Convolutional networks were proposed by LeCun in 1989 [9], but their popularity has exploded since 2012 [8]. After classification problems, they were applied to solve the object detection problem: the simultaneous finding of rectangular areas containing objects in the image and classifying the objects contained in them. Initially, CNN-based solutions for object detection processed search for regions and classification of objects separately ([4] and others). Later methods that simultaneously search for regions and their classification (one-stage detectors) have achieved great success.

The one-stage detector principle for the object detection has become widespread with the advent of SSD ([15, 3]) and YOLO ([18, 19]) convolutional neural network architectures. One-stage detectors' key idea is that after a series of convolutions and size reductions, they produce a feature map. Each point of this map corresponds to a square region of the original image. Before training, several a priori sizes of desired objects are set for each cell of the feature map, called anchors. For each point of the feature map (i.e., for each square area in the original image), the ground truth annotation boxes intersecting with the anchors are considered, and the degree of intersection is calculated as *IOU* (intersection over union). The neural network learns to predict the necessary shift and resizing of the anchor (4 out-

puts per anchor), the degree of intersection (1 output per anchor), which is concerned as a confidence measure at the inference stage, indicating the presence of an object in the anchor area, and a class of the object (C output parameters where C is the number of classes). Thus, at each point of the feature map, the neural network learns to predict either $A \cdot (4 + 1 + C)$ output values, where A is the number of anchors, 4 outputs define coordinates of the bounding box ([18, 19]) or $A \cdot (4 + C)$ output values ([15, 3, 14]). In the latter case, confidence is included in the responses for all C classes so that for anchors whose areas do not correspond to objects, the responses for all C classes are small.

Simultaneous learning of location and class prediction is achieved using the loss function

$$L = L_{loc} + \lambda_{cls} \cdot L_{cls} \quad (1)$$

where L_{loc} is the loss function for location prediction errors, L_{cls} is the loss function for classification errors, λ_{cls} is the weighting factor.

The RetinaNet [14] further develops this approach proposing an improved loss function FocalLoss for L_{cls} component of the loss function, which provides more weight for more complex cases. See [14] for details. We used this implementation of object detection CNN in this work.

3. Our approach

3.1. The problem setting and the network architecture

Unlike conventional approaches, we do not separate stages of dots detection, grid restoration, and combining dots into characters. Instead, we find whole Braille symbols directly and simultaneously recognize them, using the object detection CNN described above. We assign each character a class label from 1 to 63 using formula $c = \sum_1^6 2^{i-1} p_i$ where p_i is 1 if the i-th point presents in the character and 0 otherwise.

We scale an input image to 100dpi resolution. Thus, the standard A4 page is scaled to approximately 864x1150 resolution. Braille characters are located with a horizontal spacing of about 25pt and line spacing about 40pt.

We use RetinaNet CNN architecture described in [14] with some minor modifications. The Optical Braille Recognition task differs in that it searches for a large number of approximately the same small size objects with a fixed width to height ratio. Therefore, we have simplified RetinaNet architecture to reduce the execution time, primarily NMS operations. Only one "output to class + box subnet" (see Figure 3 in [14]) was used at the layer level with feature map cells having 16x16 size. It guarantees that every Braille character is covered by at least one grid cell. We used only one anchor for each grid cell with a size close to the expected character size. These modifications have reduced calculation time by more than 5 times without substantial loss of recognition quality.

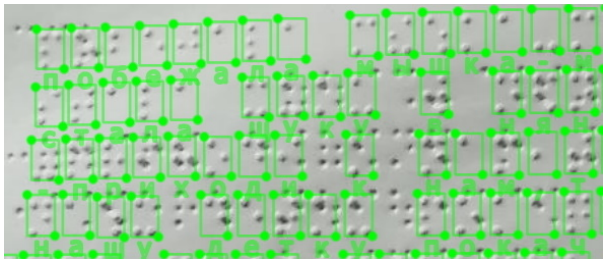


Figure 4. Sample annotations of Angelina Braille Images Dataset.

Also, we used a priory concern that Braille characters do not overlap and substantially reduced the IOU threshold used to filter overlapping detected bounding boxes using the NMS procedure. We used IOU threshold = 0.02 to allow only small overlapping due to detector imperfection.

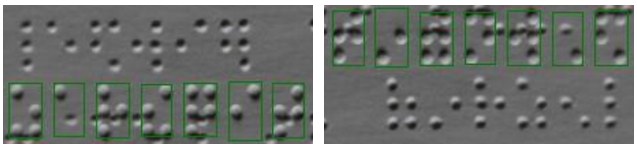
3.2. Angelina Braille Images Dataset

DSBI dataset [10] was the only publicly available dataset with labeled Braille text images, and we would like to express our deep appreciation to its creators. However, DSBI contains only scanned Braille images where Braille dots are aligned to a rectangular grid. It has a rather small number of images with limited diversity. Therefore, this dataset is not suitable for training the CNN that can properly recognize difficult images we aim to handle, neither for testing CNN designed to handle such difficult samples.

We prepared the new "Angelina Braille Images Dataset" containing 212 pages of double-sided braille books and 28 pages of student papers. These texts were photographed with various photo cameras or mobile phones under conditions close to the algorithms' intended work conditions. It includes curved pages in the open spread of the book and perspective distortions. Characters on the front side of the texts are labeled using the usual object detection problem method: for each character, a bounding box is defined, and a class from 1 to 63 is assigned, corresponding to the braille character inside the box. Sample images are shown in Figure 3, sample labels in Figure 4.

An annotation of the dataset was produced iteratively. At each iteration:

1. primary annotation was automatically generated;
2. automatically generated annotation bounding boxes were corrected manually to fit Braille text lines;
3. annotation Braille characters were converted into a plain text;



(a) (b)

Figure 5. Front side bulging dots (inside green rectangles) looks like reverse side bulged-in dots and vice versa if the image (a) is 180o rotated (b).

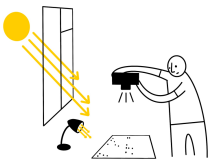


Figure 6. Light conditions Angelina Braille Images Dataset images were taken.

4. poetry texts were checked by comparison with ground truth texts of the same poems found from the Internet, considering that splitting the text into lines for poetry texts is fixed. At later iterations, non-poetry texts were checked by a spell-checker. All questionable cases were checked manually. As a result, when it is not clear if there is a Braille dot at some position or not, all ambiguous cases were resolved in favor of the grammatically correct option.

We noticed that bulging front side dots and bulged-in reverse side dots are distinguishable only when the falling light direction is known. Figure 5 shows that front side dots seems like reverse side dots and vice versa if the image is 180°rotated. We avoided this ambiguity by taking all images in light falling from the approximately top or top-left side of the page (Figure 6).

The dataset is divided into 191 training (80%) and 49 test (20%) images. Also, 44 images of various non-Braille texts from the Internet were added to the training set as negative examples.

Braille characters classes are labeled by corresponding Russian plain text letters and symbols. Still, these labels can be converted to braille dots using software tools provided with the dataset.

Our Angelina Braille Images Dataset is available at GitHub: https://github.com/IlyaOvodov/AngelinaDataset.

4. Experiments setup and results

4.1. Metrics used

We evaluated *precision*, *recall*, and *F1* metrics at a character and pixel level. Since the algorithm immediately detects symbols, it is natural to evaluate its precision at the symbol level. We used a calculation method that coincides with [12]. Detected Braille characters that intersect with ground truth characters with *IOU* (intersection over union)$\geq$ 0.5 and have a correct class are considered true positives (*TP*). Otherwise, they are considered as false positive (*FP*). So *FP* detections include both detections with either incorrect position or correct position but incorrect Braille characters assigned to them. Ground truth symbols that do not have corresponding TP detected Braille symbols are considered false negative (*FN*). The *precision*, *recall*, and *F1* metrics are defined as:

$$\begin{gathered} precision = TP/\left(TP + FP\right) \\ recall = TP/\left(TP + FN\right) \\ F1 = 2 \cdot precision \cdot recall/\left(precision + recall\right) \end{gathered} \tag{2}$$

We also evaluated per-dot metrics to compare our results with known Braille recognition methods based on dot detection. Since our algorithm does not detect individual points, we do it indirectly. All points of *TP* characters are considered as *TP* points. All points of ground truth characters that do hot have detection intersecting with them with $IOU \geq 0.5$ are considered *FN* points, and all points of detection characters that do not intersect with ground truth with $IOU \geq 0.5$ are *FP* points. For detections that intersect with ground truth characters with $IOU \geq 0.5$ but has character other than ground truth character, we compare the presence of point at each of 6 places of the g.t. and detected Braille characters. If some point present in both characters, it is considered as *TP*. Otherwise, it is considered as *FP* and *FN*, respectively. Then we evaluate Precision, Recall, and F1 at the dot level in a way described above (2).

4.2. Network training

We used the DSBI dataset [10] to compare the effectiveness of our algorithm with approaches published earlier [10, 11, 12]. Train-test split defined for DSBI dataset contains only 28 train images, which is too small for CNN training. We defined a train set as the first 74% and a test set as the last 26% pages of each Braille book in the DSBI dataset. It resulted in 84 and 30 images for train and test sets, respectively.

To evaluate our algorithm in more complex conditions, we combined the train sets from the DSBI dataset described above and our Angelina Braile Images Dataset. The evaluation was performed separately on the DSBI and Angelina Braile Images Dataset test sets.

We trained the neural network to handle images resized to 864-pixel width, which corresponds to approximately 100dpi. When training the neural network, to obtain a better resistance to different image scales and possible input distortions, train images were augmented as follows. Each

Method	Train dataset	Test dataset	Braille dot level			Character level			Perform., s/image
			Prec.	Recall	F1	Prec.	Recall	F1	
Segment [10]	DSBI [10]	DSBI	0.9172	0.9811	0.948				
Haar [10]			0.9765	0.9638	0.970				
Haar [11]			0.9838	0.9575	0.970				0.89
HOG,SVM [11]			0.9314	0.9869	0.958				15.02
SVM Grid [11]			0.9931	0.9997	0.996				1.22
TS-OBR [11]			0.9965	0.9997	0.998	0.9928	0.9996	0.9962	1.45
BraUNet [12]						0.9943	0.9988	0.9966	0.25
Our	DSBI	DSBI	0.9992	0.9995	**0.9994**	0.9977	0.9975	**0.9976**	**0.18**
	DSBI+Our		0.9984	0.9993	0.9989	0.9961	0.9964	0.9963	
	DSBI	Our	0.9812	0.9143	0.9466	0.9569	0.8980	0.9265	
	DSBI+Our	Our	0.9995	0.9986	**0.9991**	0.9985	0.9978	**0.9981**	

Table 1. Experimental results. Precesion, recall, F1 metric on a test dataset at dot and character levels, and processing performance.

image was scaled to a random width from 550 to 1150pix, which is $\pm 30\%$ of the required width. Then image was was compressed or stretched vertically by a random scale within $\pm 10\%$. Then, we rotated images at a random angle within $\pm 5°$. With a probability of 50%, we reflected the image along the vertical axis and changed each character label to the reflected character's label.

We normalized the image using the formula

$$x_c = \frac{I_c - m}{3 \max(s, 0.1 \cdot 255)} \tag{3}$$

where I_c is the intensity of the image channel in the range [0,255], c, m is the mean of I_c over the whole image, and s is the standard deviation of I_c over the image.

A random 416x416 image crop was used as CNN input. We trained the neural network for 500 epochs using Adam optimizer [7] with learning rate = 1e-4 and batch size = 24. Initially, the λ_{cls} factor in the loss function (1) was set to 1. In this case, L_{loc} component of the loss function prevails on L_{cls} resulting in more fast learning of character position than character classification. After 500 epochs, we set λ_{cls} to 100, making the contribution of both components approximately similar. We noticed that if the contribution of both components of the loss function is set equal from the very beginning, the learning process becomes unstable. Finally, we set $\lambda_{cls} = 1000$ and train the CNN using the "Reduce On Plateau" approach, i.e., reducing learning rate by factor 10 if the *F1* metric on the test set does not decrease for 500 epochs.

4.3. Results and discussion

Table 1 shows the results of the experiments.

When trained on the DSBI dataset, the new method gives *F1=0.9976* in a character-based test and *F1=0.9994* in a dot-based test. It outperforms other methods.

Investigation of characters that cause errors shows that the correct label is often questionable (Figure 7). The decision of whether detection or ground truth label is correct is, to a significant extent, subjective. Therefore a further comparison of algorithms with $F1 \geq 0.997$ on the DSBI dataset seems not informative. The reason is that the DSBI dataset was labeled without considering the semantic meanings of characters. So ambiguous cases were labeled arbitrary. The Angelina Dataset proposed in this paper was labeled using Braille characters' semantic meaning, so questionable cases were resolved in favor of an option that should present at the specific place from the grammatical point of view.

When trained on both DSBI and Angelina Braille Images Dataset, our method results in *F1=0.9981* on the Angelina test set and *F1=0.9963* on DSBI. It can be assumed that some accuracy decrease on this relatively homogeneous dataset is because CNN was trained on more diversified data, requiring more generalizing ability. As we can see, paper deformations and perspective distortion that present in photos in this dataset do not make recognition quality worse than on the DSBI dataset, where pages are flattened and scanned without distortion.

Processing of one A4 page with proposed method takes time 0.18 s/image on GPU NVIDIA 1080Ti. It outperforms other methods, including BraUNet [12], for which 0.25s/image is reported for the same hardware.

Source code and trained network weights of our algorithm are available at GitHub: https://github.com/IlyaOvodov/AngelinaReader.

Figure 7. Some characters in DSBI dataset being counted as errors. Red numbers – ground truth Braille character dots, blue – detected Braille character dots. It is not clear what one is correct.

We issued a public web service Angelina Braille Reader for recognition of Braille text images. It is available at https://angelina-reader.com.

Conclusion

In this paper, we proposed a new optical Braille recognition algorithm based on object detection convolutional neural network. The proposed algorithm has shown high accuracy and performance. This algorithm proved to be resistant to irregularities and perspective distortions of the depicted sheet with Braille text. Thus, it can recognize texts captured with a mobile phone in an everyday domestic environment.

This significantly distinguishes our algorithm from existing ones. Other methods need images with Braille text's geometric structure being undisturbed. Our algorithm's robustness makes it possible to use it as a basis for software service that can be used for the everyday needs of people who have to read Braille texts by their eyes. Such as teachers of blind students, parents of visually impaired children.

We faced that the only available dataset with annotated Braille texts does not contain enough difficult samples to evaluate Braille recognition algorithms' performance for the purposes described above. Both our algorithm and the best algorithms described before have almost 100% accuracy on this dataset. The error rate is comparable with the amount of ambiguous annotated characters.

We created and published a new Angelina Braille Images Dataset. It contains more difficult samples taken by smartphone and hand camera. Our dataset includes curved book spread pages, perspective distorted images, and other hard cases. Our algorithm has good performance on these more difficult images as well.

The proposed algorithm, the new Angelina Braille Images Dataset, and web service Angelina Reader for optical Braille recognition based on our algorithm are available on the Internet.

References

[1] Apostolos Antonacopoulos and David Bridson. A robust braille recognition system. In *International Workshop on Document Analysis Systems*, pages 533–545. Springer, 2004. 2, 3

[2] JP Dubus, M Benjelloun, V Devlaminck, F Wauquier, and P Altmayer. Image processing techniques to perform an autonomous system to translate relief braille into black-ink, called: Lectobraille. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1584–1585. IEEE, 1988. 1

[3] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017. 3, 4

[4] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 3

[5] Samer Isayed and Radwan Tahboub. A review of optical braille recognition. In *2015 2nd World Symposium on Web Applications and Networking (WSWAN)*, pages 1–6. IEEE, 2015. 2

[6] Hiroyuki Kawabe, Yuko Shimomura, Hidetaka Nambo, and Shuichi Seto. Application of deep learning to classification of braille dot for restoration of old braille books. In *International Conference on Management Science and Engineering Management*, pages 913–926. Springer, 2018. 2, 3

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 3

[9] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 3

[10] Renqiang Li, Hong Liu, Xiangdong Wang, and Yueliang Qian. Dsbi: Double-sided braille image dataset and algorithm evaluation for braille dots detection. In *Proceedings of the 2018 the 2nd International Conference on Video and Image Processing*, pages 65–69, 2018. 2, 3, 4, 5, 6

[11] Renqiang Li, Hong Liu, Xiangdong Wang, and Yueliang Qian. Effective optical braille recognition based on two-stage learning for double-sided braille image. In *Pacific Rim International Conference on Artificial Intelligence*, pages 150–163. Springer, 2019. 2, 3, 5, 6

[12] Renqiang Li, Hong Liu, Xiangdong Wang, Jianxing Xu, and Yueliang Qian. Optical braille recognition based on semantic segmentation network with auxiliary learning strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 554–555, 2020. 3, 5, 6

[13] Ting Li, Xiaoqin Zeng, and Shoujing Xu. A deep learning method for braille recognition. In *Proceedings of the 2014 International Conference on Computational Intelligence and Communication Networks*, pages 1092–1095, 2014. 3

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 3, 4

[16] GIOVANNA Morgavi and Mauro Morando. A neural network hybrid model for an optical braille recognitor. In *International Conference on Signal, Speech and Image Processing 2002 (ICOSSIP 2002)*. Citeseer, 2002. 2, 3

[17] TDSH Perera and WKIL Wanniarachchi. Optical braille recognition based on histogram of oriented gradient features and support-vector machine. *International Journal of Engineering Science*, 19192, 2018. 2

[18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3, 4

[19] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 3, 4

[20] Joko Subur, Tri Arief Sardjono, and Ronny Mardiyanto. Braille character recognition using find contour and artificial neural network. *JAVA Journal of Electrical and Electronics Engineering*, 14(1), 2016. 3

[21] V Udayashankara et al. A review on software algorithms for optical recognition of embossed braille characters. *International Journal of computer applications*, 81(3):25–35, 2013. 2

[22] Gayatri Venugopal-Wairagade. Braille recognition using a camera-enabled smartphone. *Int J Eng Manuf*, 4:32–39, 2016. 2, 3

[23] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001. 3

[24] Shanjun Zhang and Kazuyoshi Yoshino. A braille recognition system by the mobile phone with embedded camera. In *Second International Conference on Innovative Computing, Informatio and Control (ICICIC 2007)*, pages 223–223. IEEE, 2007. 2, 3

IOP Conference Series:
Materials Science and
Engineering

PAPER • **OPEN ACCESS**

Recognition of Image Pattern To Identification Of Braille Characters To Be Audio Signals For Blind Communication Tools

To cite this article: Ramiati *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **846** 012008

IOP Conf. Series: Materials Science and Engineering **846** (2020) 012008 doi:10.1088/1757-899X/846/1/012008

Recognition of Image Pattern To Identification Of Braille Characters To Be Audio Signals For Blind Communication Tools

Ramiati*, Siska Aulia, Lifwarda, Nindya Satriani S.Ningrum

Electrical Engineering, State Polytechnic of Padang, Padang, Indonesia

*ramiati76@gmail.com

Abstract. The five senses are a source of information in humans. The sense that is the main source of information is the sense of sight. Some humans are created with limited sense of sight. The blind performs reading and writing activities using Braille characters. Braille characters are writing or printed systems for the visually impaired in the form of codes consisting of six dots in various combinations that are highlighted on the paper so that they can be touched. Those with significant visual impairment need special education or learning services. Information is very important for everyone, including blind people. Submission of information is done through various media. One media that is often used is print media such as books. But books available on the market do not adapt the way for blind people to capture information. This research offers alternative solutions to overcome the above problems, namely communication aids for reading the blind. The implementation is in the form of a scanner and webcame that is equipped with a braille character text to speech system as an alternative to the lack of blind reading media, especially braille print books. The method, the reading of Braille character scripts by studying braille characters from a to z. First, a webcame or scanner captures braille characters. Second, the system will convert Braille characters and translate Braille characters into alphabetical form through Optical Character Recognition (OCR) image processing. Recognition of Braille character patterns in written text using Artificial Neural Networks (ANN). The results of research on braille character testing are in the form of alphabetical texts a to z, and audio signals of alphabet pronunciation. The results of testing the introduction of braille character patterns using a scanner for training data 100% and for testing data 90.38 % and 96.15 %.

1. Introduction

Braille character converter system is a system that can translate Braille characters into latin characters automatically. The system changes the image of braille letters and the identification of alphabetical letters and binary data from Arduino UNO. From the results of testing Braille with the camera using a time from 3 seconds to 5 seconds, the accuracy rate is 92.3% [1]. The system will recognize Braille characters and classify Braille characters into text using the Optical Character Recognition (OCR) method. Then the system will change the braille character that has become a latin character (text) into sound using a diphone system that will chop words into syllables. Diphone system used is using the Text To Speech application [2].

Inspired by the superiority of the human brain, the field of Artificial Neural Network (ANN) or a network of artificial brain cells has been developed to enable the learning process to be implemented in a system [3]. ANN is used to identify braille characters [4]. This new breakthrough makes the intelligence system a step further [3]. Braille research continues to be developed, both hardware and software. Braille research uses computer vision[5]. This research will design a braille character translator system using the Artificial Neural Network (ANN) method. Broadly speaking, this research is braille image processing with sound output with a mini PC, so that braille letters or braille character scripts can be enjoyed by listening. This literature review aims to create a Braille character translator system into an audio signal using ANN through image processing.

2. Methodology

The research method process in this research uses an experimental method which consists of the study of literature, hardware design, and software development.

2.1 Literature study

At this stage the literature discussion activity of a study is carried out by planning a braille characters recognition system in which the results are written text and sound signals, where the writer collects data and learns relevant basic theories from various sources.

The following are the stages of the research methods used in this study as follows:

1. Data Collection

In this study the data used include:

a. Sample Data

Braille characters or letters are letters used by blind people to read and write. Braille characters data with a six-point pattern in the form *.jpg.

Figure 1. Braille Characters[6]

b. Target data

This data was obtained from training data which is in the form of targets of letters of the alphabet that identify from braille characters.

c. Training data and testing data

This data is used as a training and testing process system.

2. Data Processing

In data processing, data in the form of braille character images are processed using digital image processing principles to obtain the characteristic features of each braille character. Digital signal processing to identify braille is also applied at this stage with the support of a raspberry pi device[7].

2.2 Hardware Design

At this stage the activity carried out by the writer is to make a hardware design based on Figure 2.

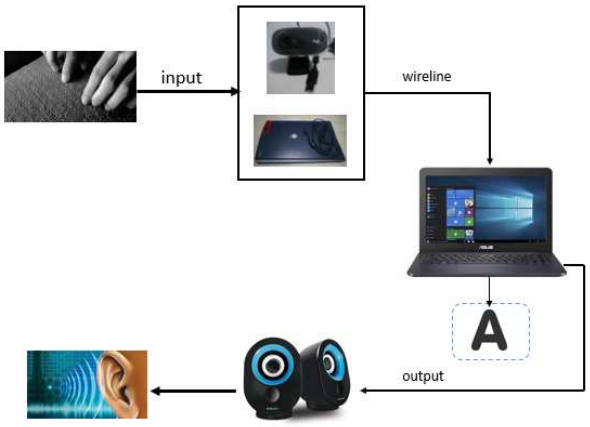


Figure 2. Design of Communication Devices for the Blind Text To Speech

At this stage, the interface display is designed based on the Graphical User Interface (GUI). The purpose of using the interface design based on this GUI is to facilitate the display in the use of the system.

In this study the manufacture of hardware or systems that are built begins with taking pictures of braille characters. The hardware used in this system is a PC (laptop), camera or webcam (Logitech C270 HD), scanner (Canoscan Lide 20) and speakers as sound output. Braille characters are captured by webcame and scanner. Communication between the laptop and webcam / scanner uses Wireline.

2.3 Software Development

At this stage the authors carry out the process of design and manufacture of software for data acquisition, image capture braille characters. Braille image processing using matlab[8][9], block diagram of braille character image processing in figure 3.

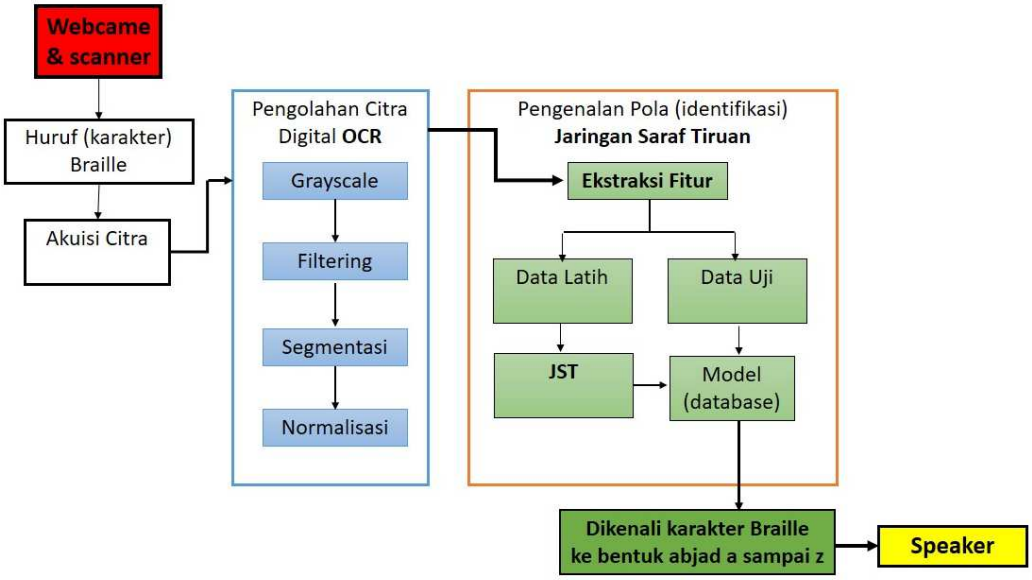


Figure 3. Braille Character Signal Processing Block Diagram

Based on Figure 3, digital image processing is the camera capturing braille letters. The result is a digital image of braille letters analyzed and image acquisition per character[10]. Then the results of image acquisition are processed using digital image OCR to identify the image of letters and numbers that are converted into writing files. The stages of digital image processing are grayscale, filtering,

segmentation and normalization. After that, the feature extraction process uses the Multi-level thresholding method in the form of a binary pattern.

Then proceed with the introduction of image patterns. The result of image feature extraction is to obtain certain characteristics of the observed letter characters. Image data that has features is saved and pattern recognition is performed. The process of pattern recognition of an image is divided into two, namely the process of training data and test data. The training data is obtained using the Artificial Neural Network method to produce a training model as a database. The training model is used for the testing process. The test results in the form of identification of braille character identification. Finally audio signal processing. This stage produces sound signals based on identification of recognizable braille letter patterns.

3. Result and Discussion

The review of literature studies in this study is limited to journals published in 2010 to 2018. The results of the implementation of the introduction of braille character patterns into sound signals using matlab. Image processing and pattern recognition of braille images using Artificial Neural Networks and output into voice signals using text to speech. Pattern recognition consists of training and testing. Braille letters are taken from braille books that are captured using a *webcam (Logitech C270 HD Webcame)* and scanner *(Scanner Canon Canoscan Lide 20*) with an image resolution of 165 pixels x 110 pixels.

At the training stage using the Radial Basis Function Network (RBFN) Neural Network. The accuracy of training and testing depends on the training parameters. The training used 48 samples of braille font images where 26 samples used HVS paper and 26 samples used drawing paper. Table 1 below shows the training parameter values.

Table 1. Training Parameters

No	Parameter	Value
1	Max. number of neurons	15
2	Number of neuron to add between display	25
3	Param. goal	1e-06
4	spread	1

- Braille Character Image Processing Results

At this stage, the data set is processed using digital image processing techniques to obtain feature extraction from each braille character. The process begins with enhancement techniques, namely the transformation of RGB (color) images into grayscale form. The goal is to determine the distinguishing regions in the image with the condition of only two black and white colors. Figure 4 Data Processing Results of Braille Character Images.

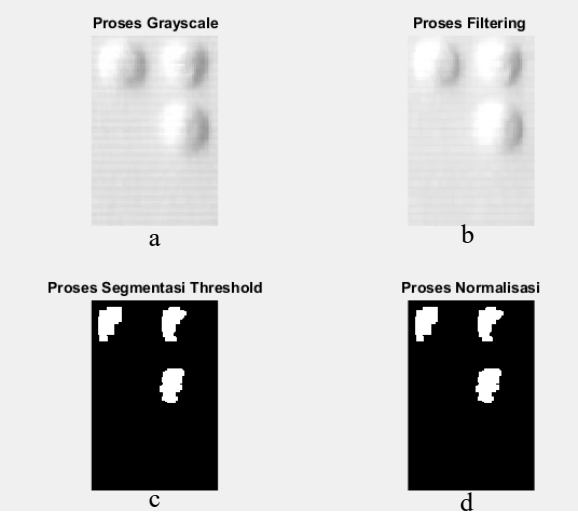


Figure 4. Data Processing Results of Braille Character Images

The second process is filtering to reduce noise, so the resulting image is smoother. Filtering results can be seen in Figure 4b. The third process of segmentation is to divide the image into segments and focus on specific objects. The results of segmentation like Figure 4c. And the final image processing is normalization. Normalization is done before entering the recognition process. This process aims to adjust the input image data with image data in the database. The results of normalization are shown in Figure 4d.

The result of the extraction of the braille letter images. Feature extraction aims to determine the characteristics of each of each braille character. The feature extraction method used in this study is a binary pattern. Multi-level thresholding is an image segmentation method that uses two or more threshold values. Its characteristic extraction technique is converting pixel values to grayscale images into binary or Multi-level thresholding. The results of the extraction of braille characters can be seen in table 2 and figure 5.

Table 2. Data Extraction Results Braille Characteristics

Feature1	Feature2	Feature3	Feature 4	Feature5	Feature6	Target
1	0	0	0	0	0	a
1	0	1	0	0	0	b
1	1	0	0	0	0	c
1	1	0	1	0	0	d
1	0	0	1	0	0	e
1	1	1	0	0	0	f
1	1	1	1	0	0	g
1	0	1	1	0	0	h
0	1	1	0	0	0	i
0	1	1	1	0	0	j
1	0	1	0	0	0	k
1	1	1	0	0	0	l
1	1	0	0	1	0	m
1	1	0	1	1	0	n
1	0	0	1	1	0	o
1	1	1	0	1	0	p
1	1	1	1	1	0	q
1	0	1	1	1	0	r

0	1	1	0	1	0	s
0	1	1	1	1	0	t
1	0	0	0	1	1	u
1	0	1	0	1	1	v
0	1	1	1	0	1	w
1	1	0	0	1	1	x
1	1	0	1	1	1	y
1	0	0	1	1	1	z

The results of identification of braille letter pattern recognition for 100% training data and 100% testing data, can be seen in table 3. For experiments on braille letter pattern recognition can be seen in test 1 (HVS paper) with 90.38 % accuracy results. Testing 2 (drawing paper) with an accuracy of 96.15 %. Testing 1 and testing 2 using a scanner, the results of its implementation can be seen in Figure 7. Testing 3 using a webcam with an accuracy of 57.69 %, can be seen in figure 6. The results of Braille Recognition Experiment can be seen in table 3.

Table 3. Result of Braille Recognition Experiment

Phase	Percentage	Error
Training	100 %	0
Testing	100 %	0
Testing 1	90.38 %	9.62 %
Testing 2	96.15 %	3.85 %
Testing 3	57.69 %	42.31%

In the next stage, an interface is implemented using the Matlab Interface GUI. Figure 5 shows the results of the implementation of the interface in this study. Figure 6 and Figure 7 are the experimental display of braille character identification experiments.

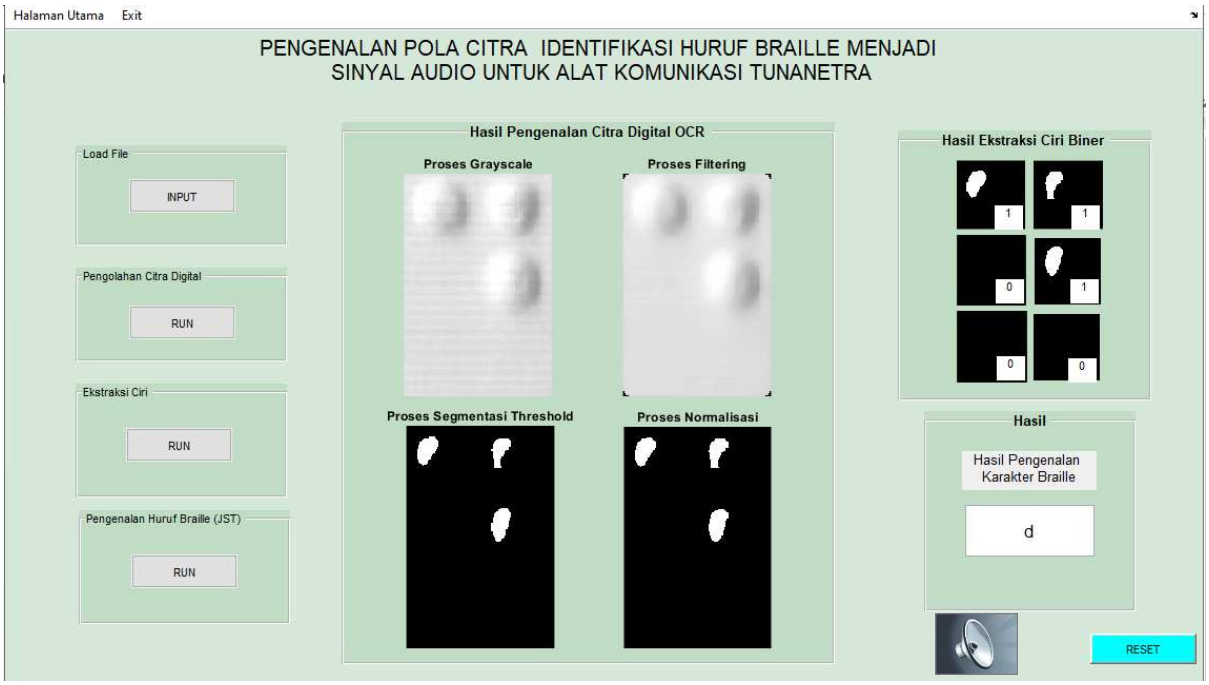


Figure 5. Results of the Implementation of the Pattern Recognition System for Braille Character Images Into Sound Signals

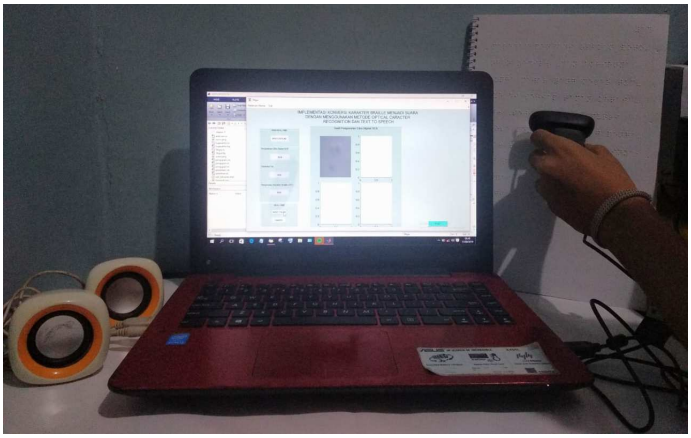


Figure 6. Integration Design of Real Time Braille Character Identification Program using Webcam

Braille character test results using a webcam are far from perfect because of its low accuracy. The system encountered an identification error of 42.31%. Errors due to several reasons, namely the object or sample that is too small, the webcam's focus in recognizing the sample, the webcam's distance from the sample, and the intensity of the light used. The test results can be seen in Figure 6.

Figure 7. Display Experiment Identification of Braille Letters into Audio Signals for Blind Communication Tools using Scanners

4. Conclusion

Based on testing and evaluation results of this study can be concluded as follows:

1. Artificial Neural Network pattern recognition method with feature extraction of multi-level thresholding binary patterns can recognize Braille characters well with testing accuracy of 100%, 90.38%, 96.15% and 57.69%.
2. This research can be a reference to facilitate communication with the blind. The implementation in this study uses matlab integration with the device, while testing in realtime in the process of using the Raspberry Pi mini PC.

References

[1] D. P. Sari and S. Rasyad, “Identifikasi Huruf Braille Berbasis Image Processing Secara Real Time,” no. November, pp. 1–2, 2017.
[2] L. P. Eka Damayanthi, “Pengembangan Aplikasi Text To Speech Dalam Pembuatan Kamus

Untuk Tunanetra," *J. Pendidik. Teknol. dan Kejuru.*, vol. 11, no. 1, pp. 1–10, 2014.
[3] Y. Eninggar *et al.*, "Pengenalan Huruf Braille Berbasis Jaringan Syaraf Tiruan Metode Hebbrule," pp. 1–5.
[4] K. Smelyakov and A. Sakhon, "Braille Character Recognition Based on Neural Networks," *2018 IEEE Second Int. Conf. Data Stream Min. Process.*, pp. 509–513, 2018.
[5] M. P. Arakeri, N. S. Keerthana, M. Madhura, A. Sankar, and T. Munnavar, "Assistive Technology for the Visually Impaired Using Computer Vision," *2018 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2018*, pp. 1725–1730, 2018.
[6] S. Hossain, A. A. Raied, A. Rahman, and Z. R. Abdullah, "Text to Braille Scanner with Ultra Low Cost Refreshable Braille Display," *2018 IEEE Glob. Humanit. Technol. Conf.*, pp. 1–6, 2018.
[7] E. Ronando and A. Sudaryanto, "Sistem Pengenalan Pola Huruf Braille Berbasis Audio Menggunakan Metode Naïve Bayes," vol. 3, no. 1, pp. 42–51, 2018.
[8] V. V. Murthy and M. Hanumanthappa, "Improving Optical Braille Recognition in Pre-processing Stage," *ICSNS 2018 - Proc. IEEE Int. Conf. Soft-Computing Netw. Secur.*, pp. 1–3, 2018.
[9] M. Wajid and M. W. Abdullah, "Imprinted Braille-Character Pattern Recognition using Image Processing Techniques," no. Iciip, pp. 0–4, 2011.
[10] S. Isayed and R. Tahboub, "A Review of Optical Braille Recognition," *2015 2nd World Symp. Web Appl. Networking, WSWAN 2015*, pp. 1–6, 2015.

Recognition of Double Sided Amharic Braille Documents

Article *in* International Journal of Image, Graphics and Signal Processing · April 2017

DOI: 10.5815/ijigsp.2017.04.01

2 authors:

Hassen Seid Ali
Kifiya Financial Technology PLC
1 PUBLICATION **7** CITATIONS
SEE PROFILE

Yaregal Assabie
Addis Ababa University
56 PUBLICATIONS **344** CITATIONS
SEE PROFILE

I.J. Image, Graphics and Signal Processing, **2017, 4, 1-9**
Published Online April 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2017.04.01

Recognition of Double Sided Amharic Braille Documents

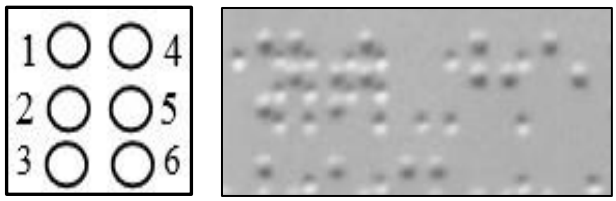
Hassen Seid Ali and Yaregal Assabie
Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia
Email: {hssnseid@gmail.com, yaregal.assabie@aau.edu.et}

***Abstract*—**Amharic Braille image recognition into a print text is not an easy task because Amharic language has large number of characters requiring corresponding representations in the Braille system. In this paper, we propose a system for recognition of double sided Amharic Braille documents which needs identification of recto, verso and overlapping dots. We use direction field tensor for preprocessing and segmentation of dots from the background. Gradient field is used to identify a dot as recto or verso dots. Overlapping dots are identified using Braille dot attributes (centroid and area). After identification, the dots are grouped into recto and verso pages. Then, we design Braille cell encoding and Braille code translation algorithms to encode dots into a Braille code and Braille codes into a print text, respectively. With the purpose of using the same Braille cell encoding and Braille code translation algorithm, recto page is mirrored about a vertical symmetric line. Moreover, we use the concept of reflection to reverse wrongly scanned Braille documents automatically. The performance of the system is evaluated and we achieve an average dot identification accuracy of 99.3% and translation accuracy of 95.6%.

***Index Terms*—**Amharic Braille Recognition, Direction Field Tensor, Double Sided Braille, Recto Dot and Verso Dot Identification, Braille Code Translation

I. INTRODUCTION

Braille is a tactile format of written communication for people with low vision and blindness worldwide since its inception in 1829 by Louis Braille [1]. It is a system of writing that uses patterns of raised dots to inscribe characters on paper [2]. A Braille cell consists of 6 dots, 2 across and 3 down, which is considered as the basic unit for all Braille symbols as shown in Fig. 1(a). The 6 dots totally give 2^6 = 64 different possible combinations of Braille character. This represents a single character, for example in English and Arabic languages [2, 3, 4]. Although most countries adopt and define their Braille code so as to fit into their local language characters, Braille systems used in the world currently are categorized into two levels as Grade 1 Braille and Grade 2 Braille [3]. Grade 1 Braille is a form of braille in which print characters are represented by one Braille cell using a mode indicator character [3]. The mode indicator determines how the character is to be read. For instance, in the English Braille, the lower case letters 'a-z' and the major punctuation symbols are represented by a single braille character or Braille cell, but others such as upper case letters, digits, and italics are represented with 'shift' character as an indicator [5]. Grade 2 Braille is introduced as a result of the rigorous attempt to minimize the volume of Braille documents by contracting words so as to minimize the time required to read a Braille document as compared to Grade 1 Braille document [3]. In Grade 2 Braille, context sensitive rules which are apparently language dependent and frequently used letter groups are used for the contraction of words. These rules determine the correspondence between one or more Braille cells and the print characters. For example, in Standard English Braille, a Braille symbol may stand for 'dis' referring the word distance when it comes at the beginning of a word and 'dd' referring the word ladder when it comes at the middle of a word [5]. A Braille document can be embossed not only on a single side of the Braille document but also on both sides to overcome space consumption by single sided Braille documents. On double sided Braille, the embossing process is done with slight diagonal offset to prevent recto and verso dots interference [6, 7] as shown in Fig. 1(b).



(a) (b)

Fig.1. (a) Braille cell, (b) double sided Braille document.

Although the invention of Braille writing system is a big stride for visually impaired people, it still limits the communication to be only among them. As a result, an optical Braille recognition (OBR) system is introduced to convert Braille text into print text. Accordingly, several OBR systems have been developed over the years for various languages [1, 2, 3, 8]. In general, Braille image recognition systems have four major processes: *image acquisition and preprocessing*, *segmentation*, *feature extraction* and *recognition* [9]. Braille document images are acquired by digital devices like scanner and this is usually followed by preprocessing with the purpose of enhancing the quality of the Braille image [2, 9]. Segmentation involves identification and separation of

embossed Braille cells from the background [10]. During feature extraction phase, salient features of the Braille cells that would be used for recognition are computed [11, 12]. Finally, recognition is done by matching a set of features of Braille cells against each record in a translation table. In general, recognition of double sided Braille documents is more difficult than that of single sided as the identification of recto and verso dots in double sided Braille documents brings in additional tasks in each phase of the recognition system [7].

Braille writing system was introduced to Ethiopia in 1923 and a number of Amharic Braille documents have been produced since then [13]. With an effort to overcome space consumption, many of the Amharic Braille documents are embossed in both sides of the document. Thus, this paper presents the development of recognition system for double sided Amharic Braille documents. The remaining part of the paper is organized as follows. Section II presents related works in the area of optical Braille recognition. In Section III, we present the characteristics of Amharic Braille system. The proposed recognition system for double sided Amharic Braille documents is presented in Section IV. Section V presents experimental results, and conclusion and future works are highlighted in Section VI.

II. Related Work

Research and development on recognition of Braille documents has been reported since the 1980s [14]. Although Braille dots have similar embossing across different scripts, the way they are encoded to form equivalent print characters vary from one script to another. For example, a single cell containing six dots represents a single Latin character whereas two or more cells may be required to represent a single character in other scripts depending on the number of characters used by the scripts. This leads to the necessity of designing specific methods for recognition of Braille documents embossed for each script. As a result, various Braille document recognition methods have been proposed over the years for different languages.

One of the earliest works on recognition of Braille documents was conducted in 1985 by Fran ̧ois and Calders [14]. The system was designed to convert Braille characters into their equivalent Latin characters with relatively complex constraints on the setup of the camera and lighting. Other attempts were made during the 1990s with the aim of reducing the constraints [2, 5, 15]. Ritchings *et al.* [7] made significant improvement by designing a system that uses a commercially available flatbed scanner to acquire a grayscale image of a Braille document. Furthermore, they considered the recognition of double sided Braille documents. More recent works on recognition of Latin Braille characters focus on improving the overall performance of the system and it has now reached an accuracy of over 99% [6, 16].

Attempts were also made to recognize Braille documents into non-Latin texts. Al-Salman *et al.* [3] conducted a research on recognition of double sided Arabic Braille document. In this work, Braille images scanned using flatbed scanner were converted to grayscale and image thresholding algorithm was developed to examine the value of each pixel so as to classify a pixel into one of the three classes (dark, light and background) based on variation of intensity level. Using two threshold values (*low* and *high*, calculated from grayscale values of the image), dot detection algorithm was developed to classify a dot as recto (contains bright region at the top and dark at the bottom) or verso (contains dark region at the top and bright at the bottom). Finally, recognition was made by translating each Braille cell into its corresponding Arabic character. Furthermore, Al-Shamma and Fathi [17] developed a system that transcribes the results of Arabic Braille recognition into text and voice.

Jiang *et al.* [8] developed a system that segments Mandarin Braille words and transforms to Chinese characters. Word segmentation consists of rule, sign and knowledge bases for disambiguation and mistake correction using adjacent constraints and bi-directional maximal matching, and segmentation precision is reported to be better than 99%. The overall translation accuracy of Mandarin Braille codes to Chinese characters for common documents was 94.38%. On the other hand, several languages of India use a largely unified braille script for writing known as Bharati Braille [18]. Bharati Braille recognition has recently received attention from researchers and developers. Accordingly, a database of Bharati Braille document image was developed to assist the effort in the development of Indian Braille recognition system [18].

Hassen and Assabie [1] conducted a research on recognition of Amharic Braille characters in single sided documents using direction field tensor [19] for noise removal and isolation of braille dots from the background. A half character detection method, which differentiates braille dots from noise, was applied and braille cells were formulated by way of examining horizontal distances between half characters. The system was also designed to determine braille dot sizes automatically, which would enable the recognition system to be resilient to differences in the size of braille dots. It was reported that the system achieved an average accuracy of 98.5% for single sided documents.

In general, preprocessing and segmentation tasks in optical Braille recognition systems may not depend on the type of scripts (or languages) as the embossing technology is the same across various scripts. However, feature extraction and recognition steps rely on the types of scripts as the number of Braille cells representing a character in natural languages varies from one script to another. Accordingly, research and development conducted on recognition of other Braille documents cannot be directly applied for recognition of Amharic Braille documents because the techniques and algorithms used to encode Braille cell into Braille code and translation of the code into print text are different. Furthermore, to our best knowledge, there is no published work on the recognition of double sided Amharic Braille

documents so far. Thus, this work focuses on the recognition of double sided Amharic Braille documents where we have dealt with the unique characteristics of Amharic Braille cell representation embossed on both sides of documents.

III. Amharic Braille System

Amharic is the working language of the federal government and several states of Ethiopia which is currently estimated to have a population of about 100 million. Even though many languages are spoken in Ethiopia, Amharic is the *lingua franca* since it is spoken as a mother tongue by a large segment of the population and it is the most commonly learned second language throughout the country [20]. The present writing system of the language is derived from Geez. The alphabet consists of 34 base characters and six other orders making it a total of 238 characters, where a character represents syllable combination of a consonant and a vowel. For example, the base character ማ(mä) has the following six other orders: ሙ(mu), ሚ(mi), ማ(ma), ሜ(me), ም(mə) and ሞ(mo). In addition, Amharic language contains 44 labialization characters and punctuation marks. The language uses both Geez and Hindu-Arabic numerals to represent numbers.

The Amharic Braille system is organized into four fundamental groups: *consonants*, *vowels*, *numbers* and *punctuation marks*. Braille codes for consonants and vowels are represented each by a single Braille cell. Each character with syllable combination is derived from consonant and vowel, requiring two Braille cells. For instance, the character 'ሀ' (hä) is a syllable combination of consonant 'ሀ' (h) and vowel 'ä' (ä). Thus, its Braille code is derived from '1:2:5' and '2:6' which are Braille codes of 'ሀ' and 'ä', respectively as shown in Fig. 2 [21].

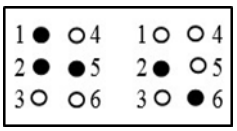


Fig.2. Braille code for character 'ሀ' (hä).

In Amharic Braille system, the same Braille code may also be used to represent three print characters (Geez numbers, Hindu-Arabic numbers or Amharic characters) as shown in Table 1. In such a case, mode indicators are used before numbers whereas, in the absence of mode indicators, Braille cells are normally interpreted as Amharic characters. The mode indicator is a Braille cell that tells Braille readers to recognize if the next Braille cell is Geez or Hindu-Arabic number. The mode indicators are represented by Braille codes of '1:2:3:4:5:6' and '3:4:5:6' for Geez and Hindu-Arabic numeral systems, respectively [21]. Punctuation marks in Amharic Braille system are represented by a combination of up to three Braille cells.

Table 1. Braille codes representing multiple print characters.

Braille Code	Symbols Represented
1	1, ፩ and ሀ
2:4	9, ፱ and ሉ
1:2	2, ፪ and ለ
1:2:4	6, ፮ and ፍ
1:2:4:5	7, ፯ and ግ

IV. The Proposed System

The proposed system architecture has six basic components: *Braille image acquisition and preprocessing*, *Braille dot segmentation*, *Braille dot identification*, *page identification*, *page transformation* and *recognition*. The general system architecture is shown in Fig. 3.

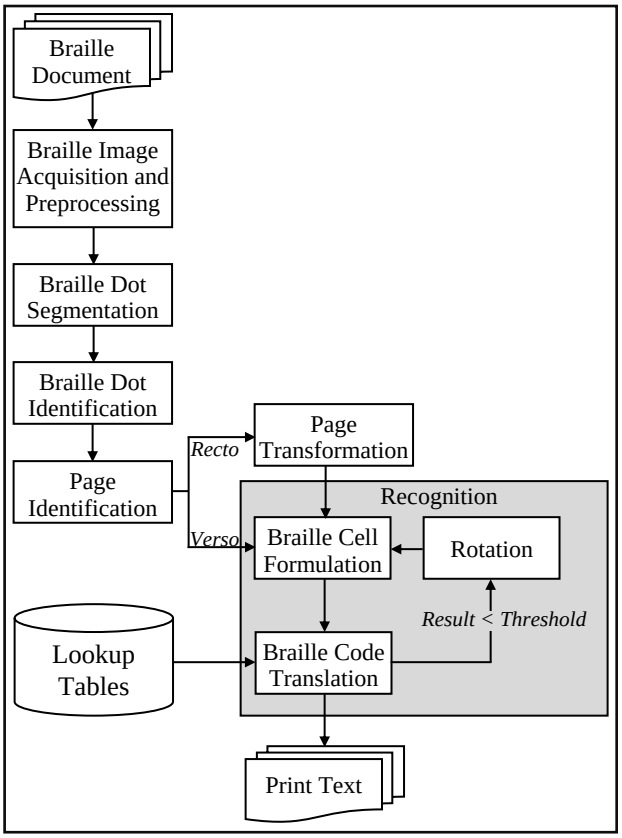


Fig.3. The proposed system architecture.

A. Braille Image Acquisition and Preprocessing

In this work, grayscale Braille images are acquired using flatbed scanner. In most low cost scanners, the document page is illuminated from an offset angle [6]. The direct implication for Braille documents is that the illumination of protrusions (recto dots) and depressions (verso dots) in that page will not be even. The face of protrusion or depression, which is angled towards the light source, will be more brightly lit and the face of protrusion or depression angled away from the light

source will be considerably less brightly lit [3, 6]. Accordingly, we exploited this property for recognition of double sided Amharic Braille documents. Fig. 4 shows grayscale image of double sided Amharic Braille document.

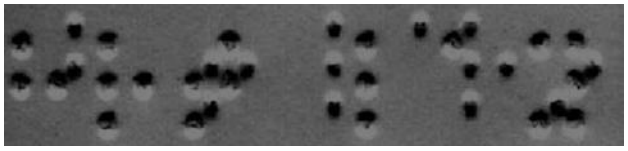


Fig.4. Double sided gray scale braille image.

After images are acquired, we employ preprocessing to make them more suitable for further processes. In this phase, our objective is to remove noise parts and highlight edges of Braille dots. To this effect, we use direction fields computed from structure tensor [19]. For a local neighborhood *f*(x,y) of an image *f*, the structure tensor *S* is computed as 2x2 symmetric matrix using Gaussian derivative operators Dx and Dy.

$$S = \begin{pmatrix} \iint (D_x f)^2 dxdy & \iint (D_x f)(D_y f) dxdy \\ \iint (D_x f)(D_y f) dxdy & \iint (D_y f)^2 dxdy \end{pmatrix} \quad (1)$$

The integrals are implemented as convolutions with a Gaussian kernel. Direction fields can be estimated from the structure tensor using complex moments which are defined as follows [19].

$$I_{mn} = \iint ((D_x + iD_y) f)^m ((D_x - iD_y) f)^n dxdy \quad (2)$$

where *m* and *n* are non-negative integers. For our preprocessing task, the order of interest to us is I_{11} which is derived from Equation (2) as follows.

$$I_{11} = \iint |(D_x + iD_y) f|^2 dxdy \quad (3)$$

I_{11} is a scalar value that measures the optimal amount of gray value changes in a local neighborhood of pixels. When integrals are implemented, due to convolutions with a Gaussian kernel, noise parts (non-linear structures) in the Braille document are suppressed. On the other hand, due to Gaussian derivative operators, I_{11} optimally highlights edges (linear structures) of Braille dots where high gray value change occurs in a local neighborhood of pixels. Fig. 5 shows I_{11} of double sided Amharic Braille image.

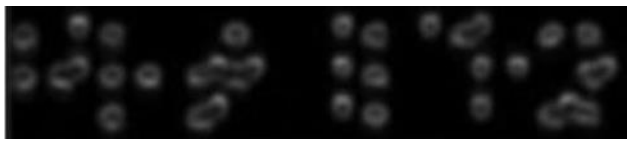


Fig.5. I_{11} of double sided Amharic braille image.

B. Braille Dot Segmentation

Braille dot segmentation separates Braille dots from the background. It is an important step in the system as recognition performance heavily depends on accurate segmentation of Braille dots. To segment Braille dots, the resultant I_{11} image is used as an input and the algorithm developed by Assabie and Bigun [22] to segment characters from the background is effectively applied for Braille dot segmentation. Fig. 6 shows the result of Braille dot segmentation.

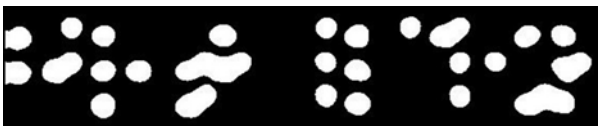


Fig.6. Braille dot segmentation

C. Braille Dot Identification

Braille dot identification classifies segmented Braille dots (isolated and overlapping) as recto and verso dots. To this end, we use dot attributes (centroid and area) and I_{10} which is derived from Equation (2) as follows.

$$I_{10} = \iint ((D_x + iD_y) f) dxdy \quad (4)$$

In fact, I_{10} corresponds to the gradient field of the image. Each pixel in I_{10} (gradient field) has magnitude and direction (angle) representing pixel intensity changes. Fig. 7 shows I_{10} of double sided Amharic Braille where colors encode the direction of pixels in the HSV color space [23] with red, yellow, green, cyan, blue and magenta colors corresponding to 0, 60, 120, 180, 240 and 300 degrees, respectively. Fig. 7 (bottom) particularly shows I_{10} (gradient field) of verso and recto dots. The direction information of pixels that form Braille dot helps to classify whether the dot is recto or verso. Empirical evidence shows that, red and cyan colors in both recto and verso dots are minimal in the gradient field.

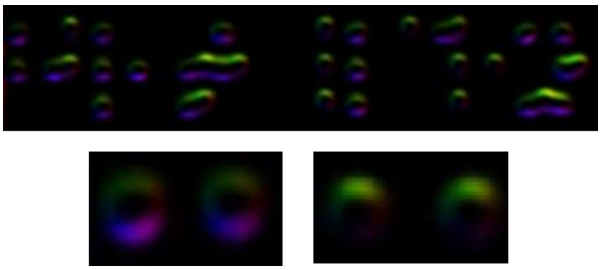


Fig.7. Top: I_{10} of double sided Amharic Braille; Bottom: I_{10} of verso dots (left) and I_{10} of recto dots (right).

However, we can see that pixels of recto dots are dominated by yellow and green colors (0-180 degrees) whereas verso dots are dominated by blue and magenta colors (180-360 degrees). This is so because of the differences in gray intensity variation observed as we go from top to bottom of each dot in the double sided grayscale Braille image (as shown in Fig. 4). It can be seen that the bright-and-dark pair appearance differs for recto and verso dots which results in different directions of pixels in the gradient field. Although I_{10} (gradient field)

is computed for the entire Braille image, for Braille dot identification, we consider only the pixels that form Braille dots. Thus, we used Algorithm 1 to group pixels that belong to Braille dots into two based on their direction.

Algorithm 1: Classification of Braille dots based on their direction.

Input: I_{10} (gradient field) of Braille image, segmented image
Output: Pixels of Braille dots categorized into two groups

```
Get Braille dot segmentation result
Get I10 (gradient field) of Braille image
If pixel is part of braille dot    //check Braille dot segmentation
    If angle is <=180        //green and yellow in I10
        PixelValue=0.5    //Group 1
    Else                    //blue and magenta
        PixelValue=1      //Group 2
    End If
Else
    PixelValue=0  //background of the Braille document
End If
```

Fig. 8 shows the grouping of Braille dots based on their direction where the gray intensities of Braille dots represent regions whose angles in the gradient field are less than or equal to 180 degrees and white parts represent Braille dot regions whose angles are greater than 180 degrees. During the process of Braille dot identification, the issue of overlapping Braille dots is a crucial concern as two or more dots can appear as physically connected with each other during Braille dot segmentation phase. The overlap may happen in the vertical or horizontal direction.

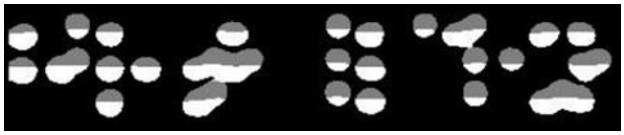


Fig.8. Braille dots extracted from gradient field.

The issue of overlap is solved by divide-and-conquer method. Vertically overlapping dots are segmented through analysis of the gradient field as the direction of pixels change from blue and magenta (180-360 degrees) to yellow and green (0-180 degrees) at the boundaries. Accordingly, in the resultant image shown in Fig. 8, we change these boundary points into background of the Braille document (thus vertically segmenting connected Braille dots) by assigning pixel values with 0 whenever a white pixel (value=1) is followed by gray pixel (value=0.5) as we go down each column. After removing vertical overlap, we develop algorithms that would recursively segment overlapping Braille dots. The algorithms are developed based on *area* and *centroid*.

Area: The area of a Braille dot corresponds to the number of pixels forming the dot. Of course, the area of Braille dot depends on the resolution used for scanning the document. With the Braille document scanned with a resolution of 200dpi, an isolated Braille dot ideally covers an average area of approximately 250 pixels with verso dots tending to have larger areas than recto dots. On the other hand, the areas covered by overlap of two, three and four dots ideally have an average area of approximately 500, 750 and 1000 pixels, respectively as depicted in Fig. 9. Accordingly, we use Table 2 to identify isolated Braille dots and estimate the number of overlapping dots.

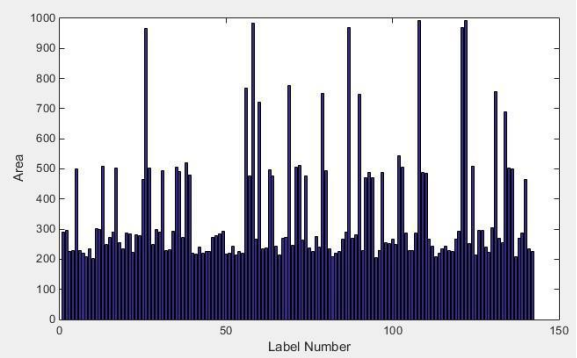


Fig.9. Areas covered by isolated and overlapping Braille dots.

Table 2. Estimation of overlaps using area

Dot Type	Area
Isolated dot	126-375
Overlap of two dots	376-625
Overlap of three dots	626- 875
Overlap of four dots	876-1125

Centroid: The centroid of a Braille dot (isolated or overlapping) locates the center of intensity distribution over the dot similar to locating the center of a mass. Once the number of overlapping dots is identified using area, the centroid provides information used to further identify the types of dots. Fig. 10 shows the position of centroids for isolated and overlapping dots.

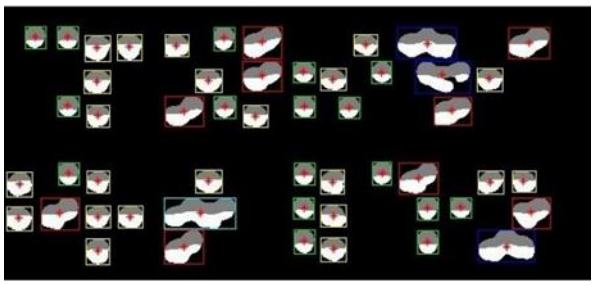


Fig.10. The position of centroids for isolated and overlapping dots.

For isolated Braille dots, the centroid of recto dots lies on the pixel whose angle is less than or equal to 180 degrees (pixel value=0.5 in Fig. 8) since recto dots are dominated by green and yellow. On the other hand, verso dots are dominated by blue and magenta (angle greater than 180 degrees) leading the centroid to lie on such colors in the gradient field. This corresponds to a pixel value of 1 in Fig. 8. However, this is not the case when two or more Braille dots are overlapping where segmentation is performed based on whether the number of overlapping dots is even or odd. When the number of

overlapping dots is odd, the centroid of overlapping dots is taken as the centroid of the middle dot. Accordingly, dots are partitioned into three regions: middle dot, left region and right region. By isolating the middle dot, it is analyzed to identify if it is recto or verso by considering the position of the centroid. When the number of overlapping dots is even, the centroid of overlapping dots is taken as the position where we segment the overlapping dots into two as left region and right region. Fig. 11 shows the arrangements of isolated and overlapping Braille dots along with their identifications in different scenarios.

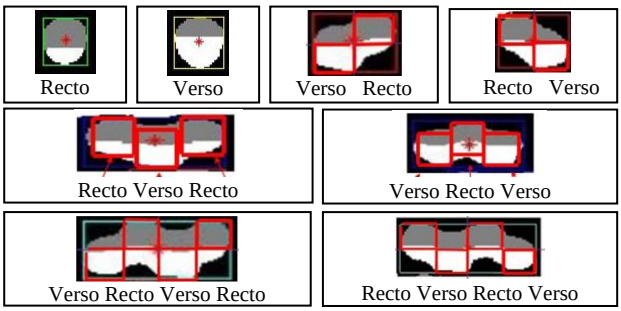


Fig.11. Arrangements of isolated and overlapping Braille dots

Left and right overlapping Braille dot regions obtained from the aforementioned processes are then recursively analyzed until all overlapping dots become isolated. Algorithm 2 shows the recursive process used to identify isolated and overlapping Braille dots. Examples of the segmentation and Braille dot identification results are shown in Fig. 12.

Algorithm 2: Procedure for Braille dot identification.

Input: Braille dot regions (along with areas and centroids)
Output: Identified Braille dots (recto and verso dots)

1. Get overlapping Braille dot region
2. Compute number of overlapping dots → **n**
3. Compute centroid
4. **If n**==1 *// if the dot is isolated*
5. **If** pixel at centroid==1
6. Dot is verso
7. **Else** *// if pixel at the centroid ==0.5*
8. Dot is recto
9. **End If**
10. **Else**
11. **If** n is odd
12. Horizontally partition the overlapping dot into **n** equally distributed regions where we have **(n-1)/2** overlapping dots to the left and another **(n-1)/2** overlapping dots to the right of the middle dot
13. Go to Line #3 to identify the middle dot
14. Go to Line #1 to identify the left region
15. Go to Line #1 to identify the right region
16. **Else** *// if n is even*
17. Horizontally partition the overlapping dot into **n** equally distributed regions where we have **n/2** overlapping dots to the left and another **n/2** overlapping dots to the right of the centroid
18. Go to Line #1 to identify the left region
19. Go to Line #1 to identify the right region
20. **End if**
21. **End If**

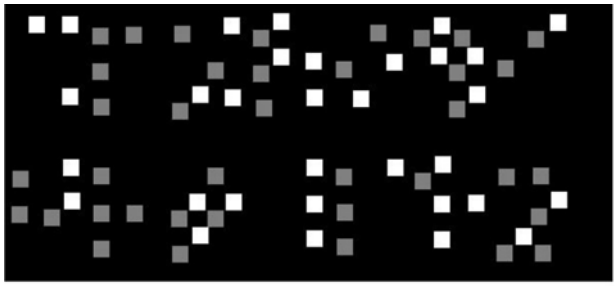


Fig.12. Identified recto and verso dots

D. *Page Identification*

For further processing, recto and verso dots which are found on the same page need to be separated in two different pages: recto and verso pages. Recto page contains only recto dots and verso page contains only verso dots. Thus, by considering the differences in pixel values of recto and verso dots, we identify separate recto and verso pages. Fig. 13 shows separated recto and verso pages constructed from the image shown Fig. 12.

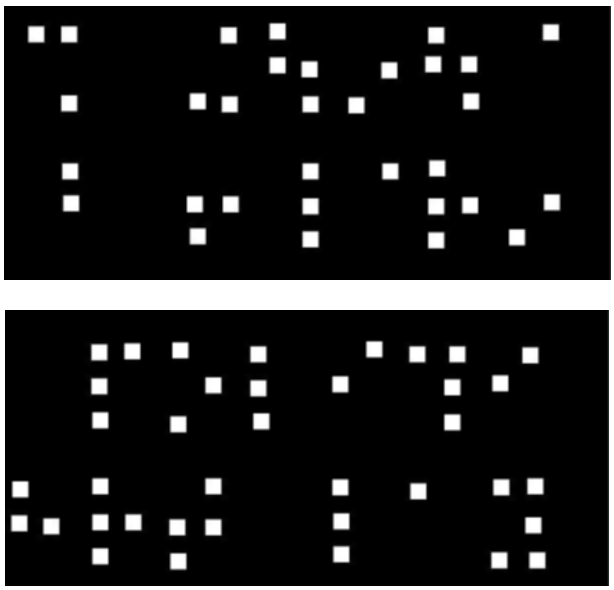


Fig.13. Separated recto (top) and verso (bottom) page.

E. *Page Transformation*

Human beings read Braille documents by touching the protruding dots in the verso page. In the case of double sided documents, reading of recto pages is performed by turning the page so as to make them verso. A similar procedure needs to be performed in automatic recognition of double sided Braille documents. This facilitates single *Braille cell encoding* and *translation algorithms* for both pages. After Braille pages are separated into recto and verso pages, recto pages are turned from right to left before Braille cells are formulated and encoded into a Braille code. In our case, this is accomplished using reflection of the recto pages about a vertical line constructed from the rightmost column of recto pages. Fig. 14 shows the result of transformation of the recto page shown in Fig. 13.

F. Recognition

Recognition of Braille documents is carried out using *Braille cell formulation* and *Braille code translation*. In the process of Braille cell formulation, up to six Braille dots are identified to construct a Braille cell. Braille cells are then encoded into a Braille code. This is achieved using Braille dot and Braille cell attributes shown in Table 3. The expected values are identified empirically and this is valid for Braille image resolution of 200dpi where the values are to be changed proportionally with the changes in resolution.

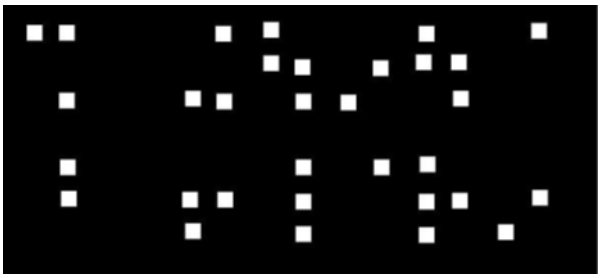


Fig.14. Braille transformation for recto page.

Table 3. Braille dot and Braille cell attributes

Attributes	Expected values in pixels
Dot width and height	12
Braille cell width	52
Braille cell height	90
Space between Braille cells	28
Space between Braille cell lines	44

It can also be seen that the horizontal and vertical spaces between Braille cells is typically greater than that of the Braille dots within the same Braille cell. Furthermore, a space amounting to a Braille cell is used to separate words in a text line. These characteristic features along with the Braille dot and cell attributes are used to identify not only Braille cells but also group Braille cells into words and text lines. After Braille cells are identified, the dots forming Braille cells are encoded into the respective six Braille codes based on their spatial positions. Fig. 15 shows the results of Braille cell identification and encoding.

After Braille cells are identified, the next step in the recognition process is to translate them into print text. Thus, Braille code translation is performed using four lookup tables that are designed based on Amharic Braille character organization. The lookup tables are organized as consonant, syllable (consonant-vowel combination), number (Geez and Arabic) and punctuation marks. The design avoids lookup tables for vowels and mode indicators because both of them are not printable symbols in Amharic language. Designing the lookup in four tables avoids ambiguity during translation because, in Amharic Braille system, different characters can be represented with the same Braille code as discussed in Section III. Subsequently, we applied Algorithm 3 to translate Braille cells into equivalent print characters.

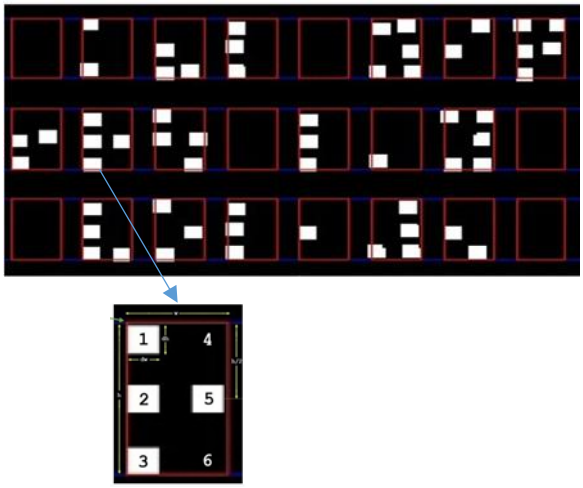


Fig.15. Braille cell identification (top) and encoding (bottom).

Algorithm 3: Procedure for translation of Braille codes.

Input: Braille code instances and lookup tables
Output: Print text

```
Get lookup tables: NUMBER, SYLLABLE, CONSONANT, PUNCTUATION
Get Braille code instance, code[i]
If code[i] is null
        Print space character
Else if code[i] is Geez number mode indicator
        Get code [i+1]
        Print the corresponding Geez number from NUMBER
Else if code[i] is Hindu-Arabic number indicator
        Go code [i+1]
        Print the corresponding Arabic number from NUMBER
Else if code[i] represents consonant
       Get code [i+1]
       If code [i+1] is vowel
              Concatenate code [i] and code [i+1] as new code
              Print the corresponding character from SYLLABLE
       Else
              Print the corresponding character from CONSONANT
       End if
Else if code[i] represents punctuation
        Print the corresponding symbol from PUNCTUATION
Else
        Code[i] is not an Amharic Braille code
End if
```

Braille recognition system is to be used by the sighted persons. However, most of them may not understand the correct layout of the Braille documents when placed on a scanner during the image acquisition phase. This may lead to turn the page upside down which results in a rotation of 180 degrees. To correct a reversed paper, we use a threshold of the translation results to determine if the page is rotated. The translation result that is considered here is not the actual recognition result but the percentage of Braille codes in a page whose corresponding entries are available in the lookup tables. Braille pages with translation results below a threshold value would be reversed by 180 degrees and Braille cells in the reversed page are reformulated for re-translation. In this experiment, a translation result of 68.5% was used as

a threshold value which was set by empirical analysis. The translation result of the reversed page is compared with the previous result and the better one is taken as the final result.

V. Experiment

To test the performance of the proposed system, we developed a prototype using MATLAB programming tool. Furthermore, we use MySQL to manage the lookup tables that are used during Braille code translation. Experiments were then conducted using real-life double sided Amharic Braille documents.

A. The Dataset

Double sided Amharic Braille documents were collected from Kennedy Library of Addis Ababa University, Ethiopian National Association for the Blind and Misrach Center. The dataset evenly contains good and bad quality documents. Braille images were then acquired using flatbed scanner with a resolution of 200dpi in "jpg" format.

B. Test Result

The performance of the proposed system architecture is measured from two perspectives: dot identification and translation accuracy. Dot identification accuracy (IA) and translation accuracy (TA) of a Braille code into print text are calculated as follows.

$$IA = \frac{(CIRD+CIVD)}{Total\ Number\ of\ Dots} * 100\% \quad (5)$$

$$TA = \frac{Correctly\ Translated\ Characters}{Total\ Number\ of\ Characters} * 100\% \quad (6)$$

where *CIRD* is Correctly Identified Recto Dots and *CIVD* is Correctly Identified Verso Dots

Accordingly, the system achieved Braille dot (recto and verso) identification accuracy of 99.3% and an average translation accuracy of 95.6%. The quality of documents had significant effect on the performance of the system. For high quality documents, the system recognized Braille documents with an accuracy of above 99%. However, old and low quality documents posed difficulty for the recognition system. Furthermore, the report on the performance of our system takes overlapping dots into consideration, which is not addressed in many of the Braille document recognition systems developed for other scripts as well.

VI. Conclusion and Future Work

Many research and development works have been conducted on recognition of documents. Braille document recognition systems are designed to recognize Braille documents (single and double sided) for different languages. Our work is the first attempt on recognition of double sided Amharic Braille documents. In this work, direction field tensor combined with gradient field show an interesting result as compared to prior works on different languages. Direction field tensor plays an important role in segmenting dots from the background and gradient field is used for identification of dots as recto and verso. This work also considers identification of overlapping recto and verso dots. The performance of the proposed system can be enhanced further by incorporating linguistic resources. A post processing component at word and sentence level is expected to enhance the performance of the recognition system. Accordingly, future work is recommended to be directed at looking into post processing activities and integration with text to speech synthesis systems.

References

[1] Miftah Hassen and Yaregal Assabie, "Recognition of Ethiopic Braille Characters," In *Proceedings of the 4th International ACM Conference on Management of Emergent Digital Eco Systems*, 2012.

[2] A. Antonacopoulos, "Automatic Reading of Braille Documents". In: H. *Bunke, P.S.P. Wang (eds.): Handbook of Character Recognition and Document Image Analysis*. World Scientific Publishing Company, pp. 703–728, 1997.

[3] A. Al-Salman, Y. AlOhali, M. AlKanhal and A. AlRajih, "An Arabic Optical Braille Recognition System," in *Proceedings of the First International Conference in Information and Communication Technology and Accessibility (ICITA 2007)*, Hammamet, Tunisia, 2007.

[4] W. Lisa, A. Waleed, and H. Stephan, "A Software Algorithm Prototype for Optical Recognition of Embossed Braille," in *the 17th conference of the International Conference in Pattern Recognition*, Cambridge, UK, August 2004.

[5] P. Blenkhorn, "A System for Converting Braille into Print," *IEEE Transactions on Rehabilitation Engineering*, vol. 3, no. 2, pp. 215-221, June 1995.

[6] A. Antonacopoulos and D. Bridson, "A Robust Braille Recognition System," In *A. Dengel and S. Marinai (Eds.), Document Analysis Systems VI*, Springer Lecture Notes in Computer Science, LNCS 3163, pp. 533-545, 2004.

[7] R. Ritchings, A. Antonacopoulos and D. Drakopoulos, "Analysis of Scanned Braille Documents", In: *A. Dengel and A. Spitz (eds.): Document Analysis Systems*, World Scientific Publishing Company, pp. 413–421, 1995.

[8] M. Jiang, X. Zhu, G. Gielen, E. Dr ábek, Y. Xia, G. Tan and T. Bao, "Braille to print translations of Chinese", *Information and Software Technology* 44 (2), 91-100, 2002.

[9] N. Falcón, C. Travieso, J. Alonso and M. Ferrer, "Image Processing Techniques for Braille Writing Recognition", *EUROCAST*, LNCS3643, pp. 379-385, 2005.

[10] A. Al-Saleh, A. El-Zaart and A. AlSalman, "Dot Detection of Optical Braille Images for Braille Cells Recognition" *ICCHP2008*, LNCS 5105, pp. 821–826, 2008.

[11] C. Ng, V. Ng, and Y. Lau, "Regular Feature Extraction for Recognition of Braille," In *Proceedings of 3rd International Conference on Computational Intelligence and Multimedia Applications*, ICCIMA'99, 1999.

[12] J. Bhattacharya and S. Majumder, "Braille Character Recognition using Generalized Feature Vector Approach", *Computer Networks and Intelligent Computing*, Springer, pp. 171-180, 2011.

[13] Berihun Girma., "The Educational Situation of the Blind," *Center for educational staff development*, Addis Ababa, EThiopia, 1994.
[14] G. Franćois and P. Calders, "The reproduction of Braille originals by means of Optical Pattern Recognition", *Proc. 5th Int. Workshop on Computerized Braille Production*, Heverlee, pp. 119-122, 1985
[15] J. Mennens, L. Tichelen, G. Francois and J. Engelen, "Optical Recognition of Braille Writing Using Standard Equipment", *IEEE Transactions of Rehabilitation Engineering*, 2(4): 207-212, 1994.
[16] M. Yousefi, M. Famouri, B. Nasihatkon, Z. Azimifar and P. Fieguth, "A robust probabilistic Braille recognition system", *International Journal on Document Analysis and Recognition* , 15: 253, 2012.
[17] S Al-Shamma and S. Fathi, "Arabic Braille Recognition and Transcription into Text and Voice", In *proceedings of 5th Cairo International Biomedical Engineering Conference* Cairo, Egypt, pp 227-231, 2010.
[18] T. Shreekanth and V. Udayashankara, "A New Research Resource for Optical Recognition of Embossed and Hand-Punched Hindi Devanagari Braille Characters: Bharati Braille Bank", *IJIGSP*, vol.7, no.6, pp.19-28, 2015.
[19] J. Bigun, T. Bigun and K. Nilsson, "Recognition by Symmetry Derivatives and the Generalized Structure Tensor," *IEEE TPAMI 26 (2)*, pp. 1590-1605, 2004.
[20] M. Lewis, G. Simons and C. Fennig, Ethnologue: Languages of the World, Seventeenth edition. Dallas, Texas: SIL International, 2013.
[21] Special Education Team, "Braille Teaching Guide (in Amharic)", Curriculum Development and Reearch Institute of Ethiopia, Addis Ababa, Ethiopia, 1998.
[22] Yaregal Assabie and J. Bigun, "Offline Handwritten Amharic Word Recognition," *Pattern Recognition Letters, 32 (2011)*, pp. 1089-1099, 2011.
[23] S. Thorstein, "CIE Div 1, R1-47 Hue angles of Elementary Colours," *International Commission on Illumination-CIE*, Norway, 2004.

Authors' Profiles

Hassen Seid Ali received Master's Degree in Computer Science from Addis Ababa University, Addis Ababa, Ethiopia. He received Bachelor Degree in Computer Science from HiLCoE School of Computer Science, Addis Ababa. He also received Bachelor Degree in Physics from Bahir Dar University, Bahir Dar, Ethiopia. His research interests are pattern recognition and digital image processing.

Yaregal Assabie received his PhD in Electrical Engineering from Chalmers University of Technology, Gothenburg, Sweden. He received Master's Degree in Information Science and Bachelor Degree in Computer Science from Addis Ababa University, Ethiopia. He is currently working as an Assistant Professor at the Department of Computer Science, Addis Ababa University. His research interests are natural language processing, pattern recognition and digital image processing.

How to cite this paper: Hassen Seid Ali, Yaregal Assabie,"Recognition of Double Sided Amharic Braille Documents", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.9, No.4, pp.1-9, 2017.DOI: 10.5815/ijigsp.2017.04.01

International Journal of Health Geographics

RESEARCH

Open Access

Comparative analysis of computer-vision and BLE technology based indoor navigation systems for people with visual impairments

Jayakanth Kunhoth[1*], AbdelGhani Karkar[1], Somaya Al-Maadeed[1] and Asma Al-Attiyah[2]

Abstract

Background: Considerable number of indoor navigation systems has been proposed to augment people with visual impairments (VI) about their surroundings. These systems leverage several technologies, such as computer-vision, Bluetooth low energy (BLE), and other techniques to estimate the position of a user in indoor areas. Computer-vision based systems use several techniques including matching pictures, classifying captured images, recognizing visual objects or visual markers. BLE based system utilizes BLE beacons attached in the indoor areas as the source of the radio frequency signal to localize the position of the user.

Methods: In this paper, we examine the performance and usability of two computer-vision based systems and BLE-based system. The first system is computer-vision based system, called CamNav that uses a trained deep learning model to recognize locations, and the second system, called QRNav, that utilizes visual markers (QR codes) to determine locations. A field test with 10 blindfolded users has been conducted while using the three navigation systems.

Results: The obtained results from navigation experiment and feedback from blindfolded users show that QRNav and CamNav system is more efficient than BLE based system in terms of accuracy and usability. The error occurred in BLE based application is more than 30% compared to computer vision based systems including CamNav and QRNav.

Conclusions: The developed navigation systems are able to provide reliable assistance for the participants during real time experiments. Some of the participants took minimal external assistance while moving through the junctions in the corridor areas. Computer vision technology demonstrated its superiority over BLE technology in assistive systems for people with visual impairments.

Keywords: Indoor navigation, People with visual impairments, Computer vision, Mobile technology

Introduction

Indoor navigation is becoming an important topic in the field of communication technology and robotics. It is the process of identifying the correct location of the user, planning a feasible path and ushering the user through the path to reach the desired destination [1]. The most important and challenging task in a navigation system is location estimation or localization of the user in real-time [2]. In outdoor environments, the location of the user is tracked using global positioning systems (GPS). GPS are not effective in indoor areas due to non-line of sight problem [3]. In order to overcome the limitations of GPS, various technologies are utilized to track the position of a user in the indoor environment [4]. Computer vision technologies [5] demonstrated their effectiveness in providing guidance for users. In the other hand, BLE-based navigation systems are still being developed and they are commercially available in the market [6]. In fact, BLE-based system are accurate but they require beacon devices to cover new areas with navigation services [7]. Based on this fact, we examine the performance and usability of a BLE-based navigation system in contrast with two computer-vision navigation systems.

*Correspondence: j.kunhoth@qu.edu.qa
[1] Department of Computer Science and Engineering, Qatar University, Al Jamiaa Street, Doha, Qatar
Full list of author information is available at the end of the article

In general, the computer-vision based systems utilize a smartphone or a device embedded with a camera such as a Google Glass to capture the scenes while the user is walking through the indoor areas. Localization using computer vision technology requires recognition of the visual scene by matching the captured picture with existing pictures or by employing trained models. The sort of these model differs depending on the targeted objectives, such as support vector machine (SVM) model [8], neural network model including deep learning model [9], and so. Apart from scene recognition some of the existing systems depend on recognizing the text or visual markers (QR codes or Barcodes) pasted in the indoor areas to provide reliable guidance for the people with VI.

BLE technology based systems make use of BLE beacons to estimate the location of the user. The BLE beacons are attached to the walls or ceilings of the indoor environments. A smartphone or a radio frequency signal receiver is used to record the RSS from the BLE beacons to estimate the position of the user. The most used position estimation techniques are lateration [10], BLE fingerprinting [11] and proximity sensing [12]. The lateration technique calculates the distance between the receiver device and the beacons. The proximity technique is based on the measurement of the proximity of the receiver to recently known locations. BLE fingerprinting applies a pattern matching procedure, where the RSS from a particular beacon will be compared with the RSS stored in the database.

In this work, we developed two computer-vision based navigation systems and a BLE based navigation system for people with VI in indoor areas. Moreover, we compared the performance of three systems which utilized three different approaches , convolution neural network (CNN) based indoor scene recognition, QR code recognition and BLE beacons based indoor positioning for guiding the people with VI in indoor areas. An indoor image data set has been built for indoor scene recognition application. We tuned the CNN model to recognize the indoor areas using the images in the dataset. This scene recognition model is extended for guiding people with VI.

All three systems are developed in a similar manner where a common android application is utilized for providing navigation aid to the users. The major difference between the three systems are the underlying positioning technique implemented on it. Moreover, except BLE based system, the rest are designed in a client-server architecture. The first computer vision based system namely 'CamNav' utilize indoor scene recognition technique to estimate the position of the user. CamNav employed a trained model to recognize the location from the query or captured images. CamNav use android application to capture the images and provide instruction to the user. The captured images are processed in the server using a trained deep learning model. The second computer-vision based system namely QRNav utilize visual markers called QR codes pasted in the indoor areas to determine the location of the user. Like CamNav, QRNav utilizes the Android application to capture the images and provide instruction to the user. The QR codes contained in the captured images are decoded by a state of art QR code decoder implemented in the server. The BLE based system consists of an Android application and an infrastructure made up of BLE beacons. Unlike computer-vision based systems, the BLE based system utilizes the Android application for users' location estimation as well as providing instructions to the users. In addition to the development of the three navigation systems, the performance and usability of the systems are analyzed in real time environment. A field study with 10 blindfolded participants was conducted while using three systems. In order to compare and contrast the performance, efficiency, and merits of the three systems we considered navigation time, errors committed by the users and feedbacks from the users about the systems for further analysis. The remaining sections of the paper are organized as follows. In "Related work" section, we detail a fundamental study about the related work. In "Systems overview" section, we give an overview of CamNav, QRNav and BLE beacon based system. In "Evaluations and results" section, we present the evaluation, experimental setup and obtained results. In "Discussion" section, we discuss our findings. And finally, in "Conclusion" section, we conclude the paper.

Related work

Considerable number of assistive systems, designed to augment visually impaired people with indoor navigation services, have been developed in the last decades. These systems utilize different types of technologies for guiding the people with VI in indoor areas. Since indoor positioning is the important task in navigation, in this segment initially we discuss various technologies and approaches utilized for positioning the user in indoor areas. Later we discuss and compare various existing navigation systems developed for people with VI.

Overview of indoor positioning approaches

Due to the inaccuracy of traditional GPS based approaches in indoor areas , high sensitive GPS and GPS pseudolites [13] are utilized for positioning the user in indoor areas. High sensitive GPS and GPS pseudolites displayed acceptable accuracy in indoor positioning, but their implementation cost is high. Apart from GPS based approaches, various technologies has

been leveraged for the development of positioning module in indoor navigation systems. Figure 1 illustrates the hierarchical classification of common indoor positioning approaches utilized in indoor navigation systems.

Computer vision-based approaches make use of traditional cameras, omnidirectional cameras, 3d cameras or inbuilt smartphone cameras to extract visual imageries from the indoor environments. Diverse image processing algorithms such as Scale Invariant Feature Transform (SIFT) [14], Speeded Up Robust Feature (SURF) [15], Gist features [16] etc. were utilized for extracting the features from the captured imageries and followed by matching the query images. Together with the feature extraction methods, conventional approaches for vision-based positioning methods utilize clustering and matching algorithms. In addition to conventional approaches, deep learning-based computer vision solutions were developed in last 5 years. Deep learning models are composed of multiple processing layers to learn features of data without an explicit feature engineering process [17]. It made deep learning based approaches distinguished among object detection and classification methods. Apart from identifying the indoor locations based on matching the images, egomotion based position estimation methods were also employed in the computer-vision based positioning approaches [18]. Egomotion is the technique of estimating the positions of camera with respect to it's surrounding environment.

Pedestrian dead reckoning (PDR) approaches estimate the position of the user based on their known past positions. PDR methods utilizes data acquired from the accelerometer, gyroscope and magnetometer to compute the position of the user. The traditional PDR algorithms compute the position of the user by integrating the step length of the user, number of steps traveled by the user and heading angle of the user [19, 20]. It is observed that conventional PDR approaches are abundant with position errors due to drift [21], varying step length of the users, sensor bias. In order to compensate the errors generated in traditional PDR approaches, most of latest PDR based systems combined another positioning technologies such as BLE or Wi-Fi along with it or introduced some novel sensor data fusion methodologies.

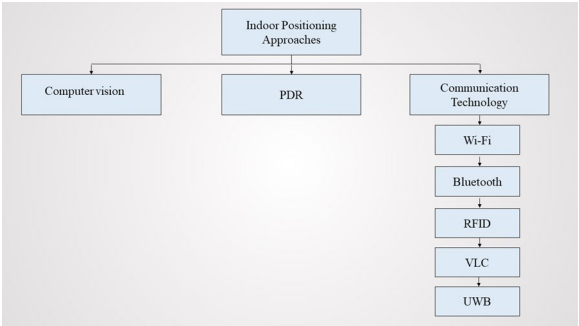


Fig. 1 Hierarchical classification of indoor positioning approaches

RFID, Wi-Fi, Ultra-Wide Band (UWB), Bluetooth and Visible Light Communication (VLC) are the popular communication technology based approaches utilized for indoor positioning task. RFID systems comprises of an RFID reader and RFID tags pasted on the objects. There exist two different types of RFID tags; active and passive. Passive tags does not require an external power supply. And many of the recent RFID based systems use passive RFID tags over active RFID tags. Time of arrival (TOA), Time difference of arrival (TDOA), Angle of arrival (AOA) and Received signal strength (RSS) the popular methods used in RFID based system for position estimation [22]. Indoor environment can contain different types of static objects such as walls, shelves etc. which can cause non-line of sight scenarios. In this context except RSS based position estimation, rest of the methods fails to compute the position of the user with minimal errors. The popular RSS based positioning approaches are trilateration and fingerprinting [23]. At present most of the indoor areas are equipped with Wi-Fi routers for providing seamless internet access for private groups or individuals or public groups. This existing Wi-Fi infrastructure can be utilized to localize the user and to provide navigation aid for the users. The Wi-Fi access points are used as the source for transmitting the signals to the receiving device (mobile or small hardware) and receiving device utilize the received signal to estimate the position of the user. Despite the fact that Wi-Fi routers are costly compared to other RF signal transmitting devices, Wi-Fi based positioning methods displayed its popularity over other methods in recent years because of the availability of Wi-Fi routers in indoor areas. Wi-Fi-based indoor positioning systems make use of RSS fingerprinting or triangulation or trilateration methods for positioning [24]. Bluetooth based systems displayed similar or better accuracy in indoor positioning while comparing with Wi-Fi based systems. They make use of Bluetooth low energy (BLE) beacons installed in indoor environments to track the positions of users using lateration or proximity sensing approaches or RSSI fingerprinting [25]. In BLE systems , RSSI fingerprinting method has displayed better positioning accuracy while comparing with all other methodologies [26]. In order to preserve the efficiency of BLE indoor positioning systems, the data from the BLE beacons should be collected within a range of 3 meters. In recent advances, mostly a smartphone is used as a receiver for both Bluetooth and Wi-Fi signals.

Existing LED and florescent lamps in the indoor areas can be utilized for developing low cost indoor positioning solutions. Nowadays these LEDs or fluorescent lamps are becoming ubiquitous in indoor environment. Visible light communication (VLC) based indoor positioning methods use the light signals emitted by fluorescent lamps or LEDs to estimate the position of the user. A smartphone camera or dedicated independent photo detector is used to detect and receive the light signals from lamps. RSS and AOA are the most popular measuring approaches used in VLC based positioning methods [27]. UWB is also a radio technology which utilizes very low energy level for short-range, high-bandwidth communications. UWB based positioning systems can provide centimeter level accuracy which is far better than Wi-Fi based, or Bluetooth based methods. UWB uses TOA, AOA, TDOA, RSS based methodology for the position estimation [28].

Table 1 illustrates comparison of indoor positioning approaches

Indoor navigation solutions for people with visual impairments

Computer vision is one of the popular technology used for the development of assistive systems for people with visual impairments. Computer vision based systems utilized two types of approach, tag based and non tag based approach to provide safe navigation for people with VI in indoor environments. Tag based system utilize some visual markers or codes attached in the indoor areas and tag or marker readers for guiding the user. Non tag based systems use the natural imageries of indoor areas or imageries of some static objects or texts found in the doors or walls in the indoor areas to guide the the people with VI in indoor areas. Moreover, 3 dimensional imageries are also utilized in the development of wayfinding or navigation system for people with VI.

Tian et al. [29] presented a navigation system for people with VI. It is composed of text recognition and door detection modules. The text recognition module employs the mean shift-based clustering for classifying the text, and Tesseract with Omni page optical character recognition (OCR) to identify the content of the text. The detection of doors is done by employing a canny edge detector. An indoor localization method has been proposed in [30]. It is based on the integration of edge detection mechanism with text recognition. Canny edge detector [31] is used to spot the edges in captured images. However, the usage of edge detection may fail in settings that have limited number of edges resulting in incorrect place recognition.

An android operating system based navigation system employ Google Glass to aid people with VI to navigate in indoor areas [32]. It uses canny edge detector and Hough line transform to detect the corners and object detection tasks. In order to estimate the distance from walls, a floor detection algorithm has been used. An indoor navigation system based on image subtraction method for spotting objects has been proposed in [33]. The algorithm of Histogram back-propagation is used for constructing the histograms of colors for detected objects. The tracking of the user is achieved by utilizing continues adaptive mean shift algorithm. A door detection method for helping people with VI to access unknown indoor areas was proposed in [34]. A miniature camera mounted on the head was used to acquire the scenes in front of the user. A computer module was included to process the captured images and provide audio feedback to the user. The door

Table 1 Comparison of indoor positioning approaches

Indoor positioning technology	Infrastructure	Hardware	Popular measurement methods	Popular techniques	Accuracy
Computer vision	Dedicated infrastructure not required	Camera or inbuilt camera of smartphone	Pattern recognition	Scene analysis	Low to medium
Motion detection	Dedicated infrastructure not required	Inertial sensor or inbuilt sensors of smartphone	Tracking	Dead reckoning	Medium
Wi-Fi	Utilize existing infrastructure of building	Wi-Fi access points and smartphone	RSS	Fingerprinting and trilateration	Low to medium
Bluetooth	Dedicated infrastructure required	BLE beacons and smartphone	RSS and Proximity	Fingerprinting and trilateration	Medium
RFID	Dedicated infrastructure required	RFID tags and RFID tag readers	RSS and proximity	Fingerprinting	Medium
VLC	Dedicated infrastructure not required	LED lights and Photo detector	RSS and AOA	Trilateration and triangulation	Medium to high
UWB	Dedicated infrastructure required	UWB tags and UWB tag reader	TOA, TDOA	Trilateration	High

detection is based on a " generic geometric door model" built on the stable edge and corner features. A computer vision module for helping blind peoples to access indoor environment was developed in [35]. An "image optimization algorithm" and a "two-layer disparity image segmentation" were used to detect the things or objects in indoor environments. The proposed approach examines the depth of information at 1 to 2 meters to guarantee the safe walking of the users.

Lee and Medioni proposed an RGB-D camera-based navigation system [36] for indoor navigation. Sparse features and dense point clouds are used to get the estimation of camera motions. In addition, a real-time Corner-based motion estimator algorithm was employed to estimate the position and orientation of objects which are in the navigation path. ISANA [37] consists of a Google tango mobile device, a smart cane with a keypad and two vibration motors to provide guided navigation for people with VI. The Google tango device is included with an RGB-D camera, a wide-angle camera, and 9 axes inertial measurement unit (IMU). The RGB-D camera was used to capture the depth data to identify the obstacles in the navigation path. The position of the user is estimated by merging visual data along with data from the IMU. Along with voice feedback, ISANA will provide haptic feedback to the user about the path and obstacles in noisy environments. The service offered by Microsoft Kinect device is extended to develop the assistive navigation system for people with VI [38]. The infrared sensor of Kinect device was integrated with RGB camera to assist the blind user. RGB camera is utilized to extract the imageries and infrared sensor is employed for getting the depth information to estimate the distance from obstacles. Corner detection algorithm was used to detect the obstacles.

A low cost assistive system for guiding blinds are proposed in [39]. The proposed system utilizes android phone and QR codes pasted in the floor to guide the user in indoor areas. 'Zxing' library was used for encoding and decoding the QR code. Zxing library for QR codes detection worked well under low light conditions. 'Ebsar' [40] utilizes a google glass, Android smartphone and QR codes attached to each location for guiding the visually impaired users. In addition to that, Ebsar utilizes the compass and accelerometer of the Android smartphone for tracking the movement of the user. The Ebsar provides instruction to the users in Arabic language. Gude et al. [41] developed an indoor navigation system for people with VI that utilizes bar code namely Semacode for identifying the location and guiding the user. The proposed system consists of two video cameras, one attached on the user's cane to detect the tags in the ground and other one attached on the glass of the user to detect tags placed above the ground level. The system provides output to the user via a tactile Braille device mounted the cane.

The digital signs (tags) based on Data Matrix 2-D bar codes are utilized to guide the people with VI [42]. Each tag encodes a 16-bit hexadecimal number. To provide robust segmentation of tags from the surroundings, the 2D matrix is embedded in a unique circle-and-square background. To read the digital signs, a tag reader which comprises a camera, lenses, IR illuminator, a computer on module was utilized. etc. of. To enhance the salience of the tags for image capture, tags were printed on 3M infrared (IR) retro-reflective material. Zeb et al. [43] used fiducial markers (AR tool kit markers) printed on a paper and attached on the ceiling of each room for guiding the people with VI. Since the markers are pasted on the ceiling, the user has to walk with a web camera facing towards the ceiling. Up on successful marker recognition, the audio associated with the recognized marker stored in the database will be played to the user. A wearable system for locating blinds utilized custom colored markers to estimate the location of the user [44]. The prototype of the system consists of a wearable glass with a camera mounted on it and a mobile phone. In order to improve the detection time of markers, markers are designed as the QR code included inside a color circle. Along with these markers, multiple micro ultrasonic sensors were included in the system to detect the obstacles in the path of the user and thereby ensuring their safety. Rahul Raj et al. [45] proposed an indoor navigation system using QR codes and smartphone. The QR code is augmented with two information. First one is the location information which provides latitude, longitude, and altitude. The second one is the web URL for downloading the floor map with respect to the obtained location information. The authors address that usage of a smartphone with a low-quality camera for capturing the image and fast movement of user can reduce the QR code detection rate.

A smart handheld device [46] utilized visual codes attached to the indoor environment and data from smartphone sensors for assisting the blinds. The smartphone camera will capture the scenes in front of the user and search for visual codes in the images. Color pattern detection using HSV and YCbCr color space method were utilized for detecting the visual code or markers in the captured image. An Augmented Reality library called "ArUco is used as an encoding technique to construct the visual markers. Rituerto et al. [47] proposed a sign based indoor navigation system for people with visual impairments. The position of the user is estimated by combining data from inertial sensors of smartphone and detected markers. Moreover, the existing signs in indoor environments were also utilized for positioning the user. ArUco marker library was used to

create the markers. The system will provide assistive information to the user via a text to speech module.

Nowadays BLE beacons based systems are becoming popular for indoor positioning and navigation application due to its low cost, and easiness to deploy as well as integrate with mobile devices. BLE beacons based systems are used in many airports, railways stations around the world for navigation and wayfinding applications. In the last 5 years, assistive systems developed for people with VI also relied on BLE technology to guide users in indoor areas. To best our knowledge there are few BLE beacons based indoor navigation systems proposed for people with VI. Most of the systems just require a smartphone and an Android or iOS application to provide reliable navigation service to the user. Lateration, RSSI fingerprinting and proximity sensing methods are the commonly used approaches for localizing the user.

NavCog [48] is a smartphone based navigation system for people with VI in indoor environment. The NavCog utilizes BLE beacons installed in the indoor environment and motion sensors of the smartphone to estimate the position of the user. RSSI fingerprint matching method was used for computing the position of user. Along with position estimation, the NavCog can provide information about the nearby point of interests, stairs, etc. to the users. StaNavi [49] is a similar system like NavCog, uses smartphone compass and BLE beacons to guide the people with VI. StaNavi was developed to operate in large railway stations. The BLE localization method utilized a proximity detection approach to compute the position of the user. A cloud-based server was also used in StaNavi to provide information about the navigation route. GuideBeacon [50] also uses smartphone compass and low-cost BLE beacons to assist the people with VI in indoor environment. GudieBeacon implemented a proximity detection algorithm to identify the nearest beacons and thereby estimate the position of the user. The system speaks out the navigation instructions using Google text to speech library. Along with audio feedback, the GuideBeacon provide haptic and tactile feedback to the user.

Bilgi et al. [51] proposed a navigation system for people with VI and hearing impairments in indoor area. The proposed prototype consists of BLE beacons attached on the ceiling of indoor areas and a smartphone. The localization algorithm uses nearness to beacons or proximity detection to compute the position of the user. Duarte et al. [52] proposed a system to guide the people with VI through public indoor areas. The system namely, SmartNav consist of a smartphone (Android application) and BLE beacons. The Google speech input API enables the user to give voice commands to the system. Android text-to-speech API is used for speech synthesis. The SmartNav utilizes multilateration approach to estimate the location of the user.

Murata et al. [53] developed a smartphone based blind localization system that utilize probabilistic localization algorithm to localize the user in a multi storied building. The proposed algorithm uses data from smartphone sensor and RSS of BLE beacons. Moreover, they introduced novel methods to monitor the integrity of localization in real world scenarios and to control the localization while traveling between floors using escalators or lift. RSS fingerprinting based localization in BLE beacons systems using fuzzy logic type 2 method displayed better precision and accuracy while compared to traditional RSS fingerprinting and other non fuzzy methods such as proximity, trilateration, centroid [54].

In addition to computer vision and BLE technology, RFID , Wi-Fi , PDR technologies are widely utilized for the development of assistive wayfinding or navigation systems for people with VI. Moreover, hybrid systems which integrate more than one technology to guide blinds or people with VI are proposed in the recent years. Table 2 illustrates Comparison of discussed indoor navigation solutions for people with visual impairments.

Many of the navigation systems discussed in this section have adopted several evaluation strategies to show their performance, effectiveness, and usability in real-time. Most of the systems considered navigation time as a parameter to show their efficiency. Some of the systems utilized the error committed by the users during navigation to asses the effectiveness of the system. The most important and commonly used approach is conducting surveys, interviews with the people who participated in the evaluation of the system. The user feedback help the authors to limit the usability issues of the proposed system.

Systems overview

CamNav

CamNav is a computer-vision based system, which utilizes a trained deep learning model to perform indoor scene recognition. Figure 2 shows an overview of the system. The architecture of the system is client-server.

Server application

The server application is responsible for processing the content of the queried images and identifying the location from the queried images. The server application utilizes a trained deep learning model to recognize the location from the image. The server application returns the location information to the mobile application in a JSON format. The server application utilizes an open source message broker namely, Active MQ [55] to communicate with the Android application.

Table 2 Comparison of discussed indoor positioning solutions for people with visual impairments

References	Technology	System	Techniques	Tested by	Test	Feedback	Accuracy	Remarks
Tian et al. [29]	Computer vision	Web camera and mini laptop	Text recognition and door detection	Blinds	Door detection and door signage recognition	Voice	Medium	(−) Motion blur and very large occlusions happen when subjects have sudden head movements
Lee and Medioni [36]	Computer vision	RGB-D camera, IMU, Laptop	Camera motion estimation	Blinds and blindfolded	Pose estimation and mobility experiments	Tactile	Medium	(−) Inconsistency in maps
Manlises et al. [33]	Computer vision	Web camera and computer	CAMShift tracking	Blinds	Object recognition, navigation time	Voice	Medium	(−) Tested with 3 blinds only
Li et al. [37]	Computer vision	Tango mobile device	Obstacle detection, camera motion estimation by combining visual and inertial data	Blindfolded	Error occurred during navigation and navigation time	Voice and Haptic	Good	Using smart cane with the system reduced the errors
Kanwal et al. [38]	Computer vision	RGB camera and Kinect sensor	Corner detection using visual and inertial data	Blind and blindfolded	Obstacle avoidance and walking	Voice	Good	(−) Infrared sensor failed under strong sunlight conditions
Al-Khalifa et al. [40]	Computer vision and motion	Google glass, Android smartphone, QR code	QR code recognition, IMU	Blinds	Error occurred during navigation and navigation time	Voice	Medium	(+) Easy to use
Legge et al. [42]	Computer vision	Digital tags, Tag reader , smartphone	Digital tag recognition	Blinds and Blind folded	Tag detection, Route finding	Voice	Medium	(+) System provided independent navigation
Zeb et al. [43]	Computer vision	Web cam, AR markers	AR marker recognition	Blinds	Normal walking	Voice	Medium	Low cost
Ahmetovic et al. [48]	BLE	Beacons and smartphone	Fingerprinting, IMU	Blinds	Normal occurred events that hinder the navigation	Voice	Medium	(−) Can't inform the users that they are in wrong way
Kim et al. [49]	BLE	Beacons and smartphone	Proximity detection, IMU	Blinds	Navigation time, task completion, deviation during navigation	Tactile and voice	Good	(+) Test was carried out in a large area (busy railway station)
Cheraghi et al. [50]	BLE	Beacons and smartphone	Proximity detection, IMU	Blinds	navigation time, Navigation distance	Haptic and voice	Medium	Improvements are required to support varying pace of walking

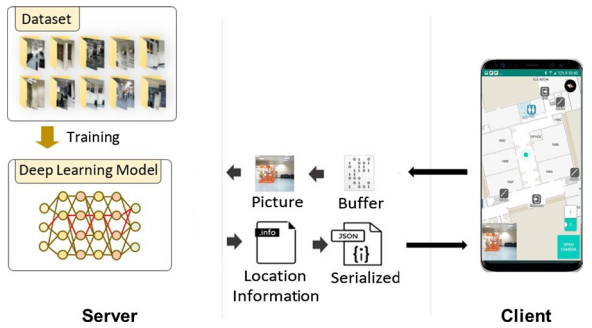


Fig. 2 CamNav system overview

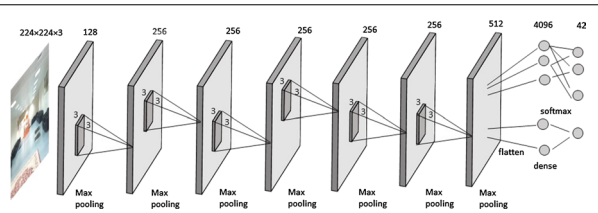


Fig. 3 Architecture of the developed deep learning model

Deep learning model

A deep learning model is configured for indoor scene recognition task. Tensorflow [56], an open-source machine learning library was utilized to build the deep learning model. Figure 3 illustrates the architecture of the developed deep learning model. The model is built using convolutional layers, pooling layers, and fully connected layers at the end.

For an input RGB image *i*, convolutional layer calculates output of the neurons which are associated to each local regions of the input. Convolutional layer can be applied to raw input data as well as output of another Convolutional layer. During convolution operation, the filter/kernel will slide over the each raw pixel of the RGB image or over the feature map generated from the previous layer. This operation compute the dot product between weights and regions of the input.

Let M_i^{l-1} be the feature map from previous layer, w_k^l is the weight matrix in current layer then convolutional operation will results new feature map M_k^l.

$$Y = \sum_{i \in N_K} M_i^{l-1} * w_k^l + b_k^l \tag{1}$$

$$M_k^l = f(Y) \tag{2}$$

Here, Y is the output of convolutional operation. N_K represents the number of kernel in current layer and b_k^l is bais value. Bais is an additional parameter used in CNN to adjust the output from the convolutional layer. Bais help the model to fit best for input data.

An activation function *f(Y)* is applied to the resulting output from the convolutional operation to generate the feature map M_k^l. Activation function a.k.a transfer function is utilized to decide the output by mapping the resulting values of convolutional operation to a specific interval such as between [0,1] or [−1,1] etc. Here we utilized Rectifier Linear Unit (ReLU) as an activation function. ReLU is the commonly used activation function in CNN and faster compared to other functions.

For an input x, ReLU function *f(x)* is,

$$f(x) = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{if } x \geq 0 \end{cases} \tag{3}$$

Once the convolutional operation is completed, pooling operation is applied on the resulting feature map to reduce the spatial size of feature maps by performing down sampling. Average pooling and max pooling are the two common functions utilized for pooling operation.

For a feature map of volume $W1 \times H1 \times D1$, pooling operation produce a feature map of reduced volume $W2 \times H2 \times D2$ where:

$$W2 = (W1 - F)/S + 1,$$
$$H2 = (H1 - F)/S + 1, \quad D2 = N_K$$

Here, S is stride and F is spatial extent. We used max pooling function in pooling operation where MAX operation is applied in a local region resulting a max value among that region.

The feature map in the form of n dimensional matrix are flattened in to vectors before feeding to fully connected layer. The fully connected layer combines the feature vector to build a model. Moreover, softmax function is used to normalize the output of fully connected layer that result the outputs representation based on probability distribution.

For an input image x, softmax function applied in the output layer computes the probability that x belongs to a class c_k by,

$$p(y = c_k|x; P) = \frac{e^{P_{c_k}^T x}}{\sum_{c_i=1}^{n} e^{P_{c_i}^T x}} \tag{4}$$

where n is the number of classes and P is the parameter of the model.

Our model consist of 7 convolutional layers where each has a max pooling layer attached to it. The first convolutional layer has 128 filters with size 3×3 and the last layer has 512 filters with size 3×3. The other convolutional layers use 256 filters with size 3×3. Max pooling layer is responsible for reducing the dimensions of the

features obtained in its preceding convolutional layer. Dimensionality reduction aids the convolutional neural network model to achieve translation invariance, reduce computation and lower the number of parameters. In the end, the architecture contains a fully connected layer with 4096 nodes followed by an output layer with softmax activation. The deep learning model was trained using more than 5000 images to identify 42 indoor location. The model takes an RGB image of size 224×224 pixels as input and classifies the image into one of the 42 class labels learned during the training phase.

Image dataset

Our indoor image dataset [57] contains more than 5000 images classified into different directories. Each directory represents one indoor location or class. Moreover, each directory contains a JSON file which contains the location information required by the Android application to locate the user. The images in the dataset are captured from the ground floor of building B09 of Qatar University. In order to consider various orientation of users, we captured pictures from different angles for the same location. The images are captured using Samsung Galaxy smartphone, LG smartphone and Lenovo smartphone. We considered the diversity of mobile phones to reserve the different sorts of pictures which are taken from varied cameras. Each images are in RGB format and reshaped into a size of 224×224 pixels. This dataset can be utilized for indoor scene recognition applications also. The sample images of the dataset are displayed in Fig. 4.

Navigation module

The navigation module is responsible for providing navigation instructions to the user. We utilized the indoor map and CAD drawings of indoor areas to create the navigation module. The navigation module contains the routing information between each point of interest to another point of interest. The navigation information inside the navigation module is stored in a JSON array format. One JSON array includes the navigation instructions for one specific route. We created the navigation instructions manually for each route.The common commands used in navigation instructions are turn right, turn left, walk straight etc. Moreover, instructions provide information such as distance between current location of the user (computed by the system) and critical locations such as junctions in indoor areas or doors or lift. The distance between each location associated with the captured visual scenes was measured manually and represented in terms of steps. For converting the distance measured in meters to step, we considered walking patterns of normal people.

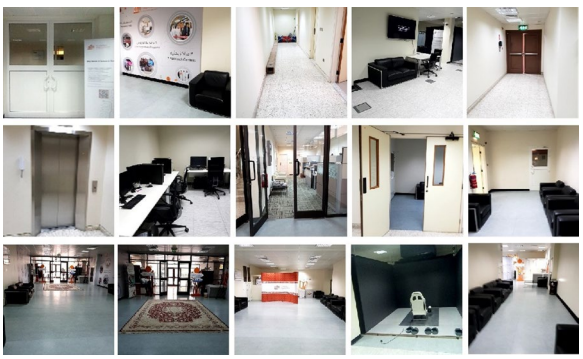


Fig. 4 Sample images from dataset

Android application

The android application captures the scenes in front of the user in a real-time manner for processing. The captured images are sent to the server application for further processing. The application contains an indoor map, a navigation module, and a speech to text module. The indoor map is created with [58]. The indoor map creation just requires computer-aided drawings of the building. The current location of the user has indicated with a turquoise color dot in the graphical interface of Android application as shown in Fig. 5. The turquoise color dot is a feature provided by Mapbox software development kit. The inputs required for the dot representation are the latitude, longitude, building number and building level. These data are made available in the directories associated with each location along with the images collected from that location in the indoor image dataset. Once the deep learning model classifies the query image, the JSON file in the directory associated with the output label will be sent to the Android application. The Android application serializes the received JSON file and updates the turquoise dot in the map. During navigation, the Android application will speak out the navigation and location information in regular intervals. A text to speech service is included inside the android application to provide voice instructions to the user. The text to speech service is developed using Android text to speech library. Android text to speech library is a popular and well accepted library used in the Android application development for the purpose of synthesizing the speech from text. Android text to speech library support many languages such as English , German, French, Japanese etc. The current version of the developed Android application can give voice instructions to the user in English language only. Visually impaired users can give instructions to the system using speech. We deployed a speech to text module in the application using Android inbuilt speech to text library for providing the mentioned service. People with

Fig. 5 GUI of Android application

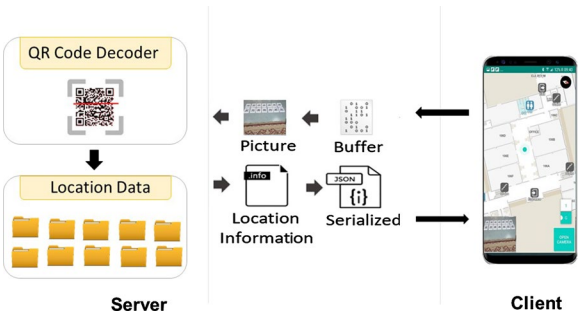


Fig. 6 Overview of the QRNav system

VI can use this service for controlling various necessary functions of the system such as to start or end navigation, to select destination, to know the current location etc. Currently system can understand English language only.

QRNav

QRNav system uses QR code recognition function to estimate the location of the user. Figure 6 depicts the overview of the QRNav system. The system consists of printed QR codes, QR code dataset, QR code decoder, and an Android mobile application.

The QR codes are attached on the ground, walls, and doors in indoor area of the building. The QR codes associated with each location have an embedded unique 4 digit id (encoded information). The android application captures the images while the user is walking. The captured images will be sent to the server. The server application contains QR code detection and decoding library to extract the unique id associated with the QR code. Once the server receives an image, the QR code detection method will look for QR codes in the image. If any QR code is detected, the decoding method will extract the unique id encoded QR code and return the location file (JSON file) associated with the unique id to Android application.

QR code decoder

We analyzed two open source barcode reader library, 'Zxing' [59] and 'Zbar'[60] for QR code encoding and decoding operations. We found that 'Zxing' library is more effective compared to 'Zbar' in low light and challenging conditions. We implemented the 'Zxing' library and the 'Zbar' library to read and extract the information from QR codes.

QR dataset

The QR dataset contains more than 25 directories where each directory is associated with an indoor location. The ame of the directories are unique (same as unique id embedded in the QR codes). The directories are enclosed with a JSON file which contains information about the location. The QR codes are created using pyzbar library (Zbar library for python language). Each indoor location is mapped with a unique QR code. The four digits unique id beginning from '1000' was manually assigned for each QR code. The QR code was printed in normal A4 sheet papers and pasted in indoor areas. In each location, we provided more than 25 copies of the QR code to provide reliable navigation service. Figure 7 shows the sample instance of attached QR codes on the floor.

Android application

The Android application contains the indoor map, text to speech module and navigation module. The text to speech module supports the speech synthesis

Fig. 7 QR codes attached on the floor

and navigation module provides navigation instruction to the user. We utilized the same Android application used in CamNav system for QRNav also.

The functioning of both CamNav and QRNav are similar. Figure 8 depicts the UML diagram of both CamNav and QRNav system.

BLE beacons based navigation

The Bluetooth low energy beacons based navigation system comprises of an Android application and BLE beacons fixed in the indoor environment. The BLE beacons based navigation system is developed with the help of indoor positioning SDK supplied by Steerpath.

BLE positioning module

The positioning module combined two popular positioning techniques to estimate the location of the user in real-time. BLE fingerprinting [11] and multilateration [61] techniques were utilized to achieve the localization of the user in the indoor map. When the system sense only two nearby beacons, then fingerprinting technique is utilized, where the observed fingerprint is compared with the pre-stored fingerprints in the database. If the system is able to detect more than two nearby beacons, the multilateration technique is used to compute the position of the user.

Beacons

Beacons: The BLE beacon infrastructure was built using the beacons supplied by Steerpath. The BLE beacons are fixed in the walls within a height of 2.5 m. Figure 9 shows the distribution of BLE beacons in the selected indoor area. The distance of separation between each beacon is maximum of 8 m.The hardware specifications of the BLE beacons are provided in Table 3.

Android application

The Android application is configured as described earlier in CamNav and QRNav to compare the three systems. However, BLE-based application does not require a server application to estimate the location of the user. The positioning module responsible for location estimation is

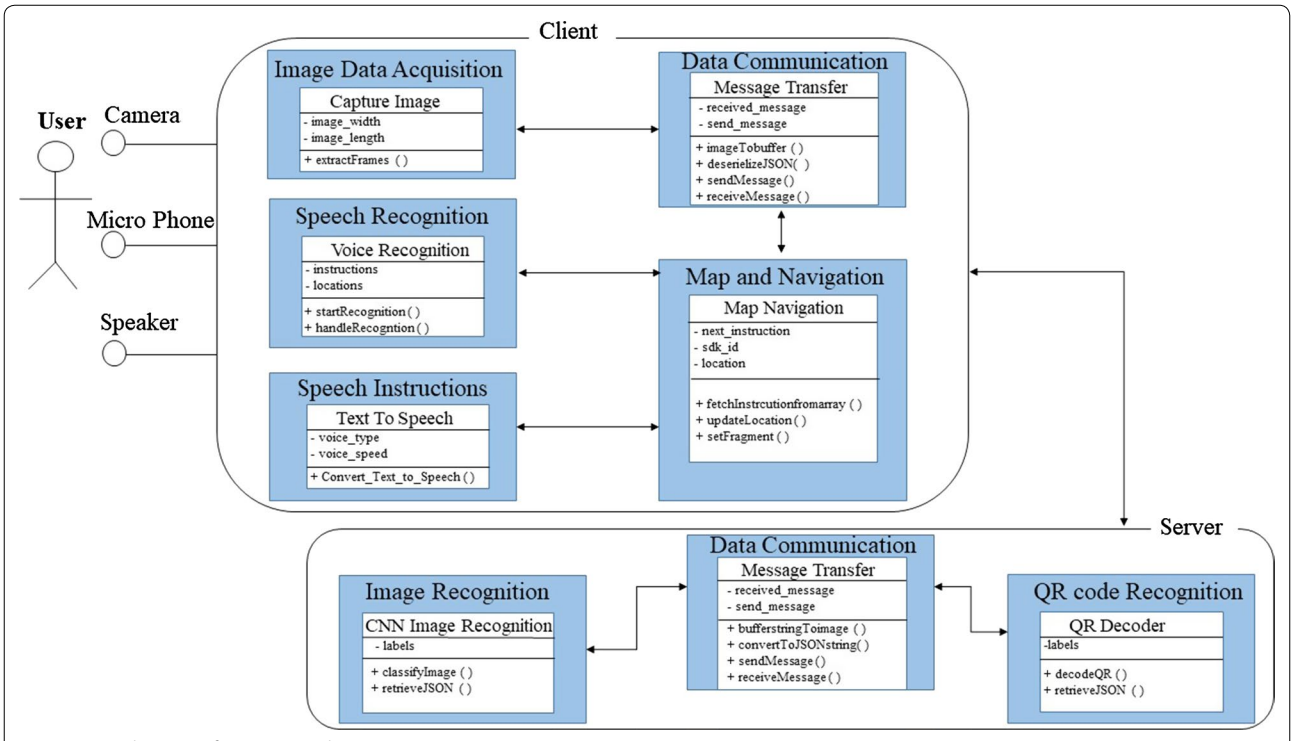


Fig. 8 UML diagram of CamNav and QRNav

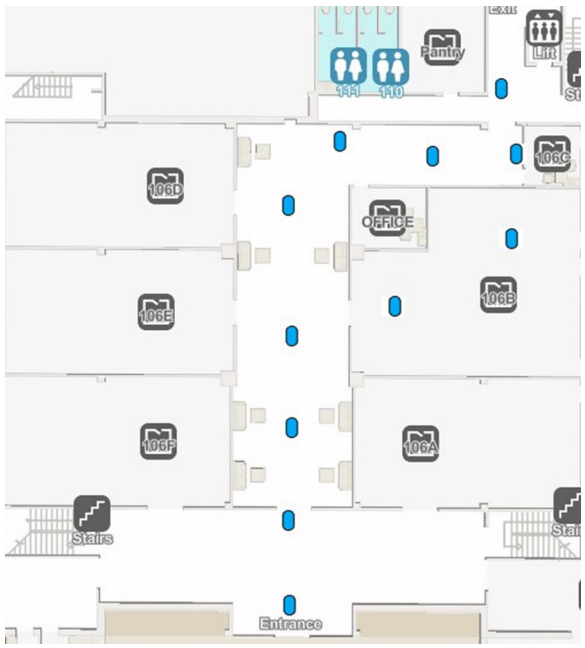


Fig. 9 Distribution of beacons over testing environment

Table 3 Hardware specification of BLE beacons

Model	Minew i3 robust smart beacon
Operating frequency	2.4 GHz (Bluetooth 4.0)
Transmission power	0 dBm output power
Transmission range	up to 50 m
Time interval	350 ms

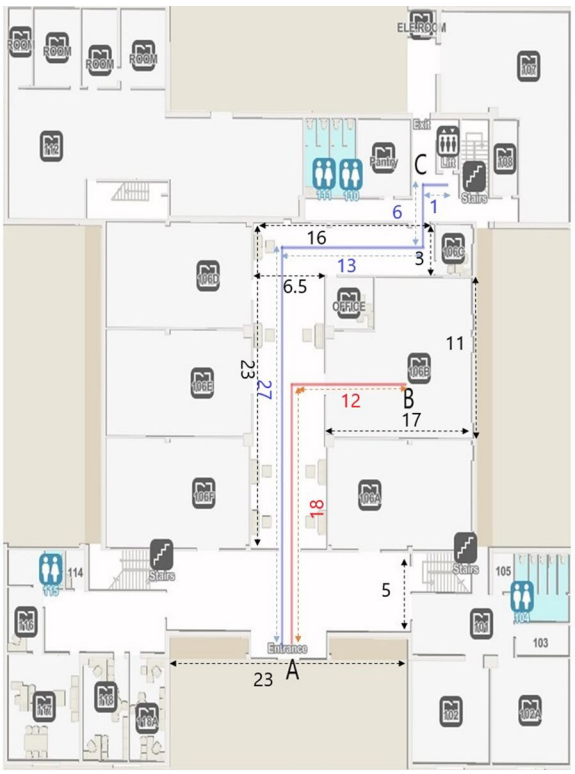


Fig. 10 Floor plan of the ground floor of building B09 (All the dimension are given in meters. Red lines and blue lines indicates the navigation routes, while black lines indicate the dimension of rooms and corridors

implemented in the Android application. The Android application receives the signal from beacons and RSSI of the signals are processed for further position estimation task. In addition to the positioning module, Android application consists of an indoor map, navigation module for providing navigation instructions and an android text to speech module for speech synthesis.

Evaluations and results

We evaluated the performance of CamNav, QRNav and BLE beacons based navigation system in a real-world environment. This study was approved by the Qatar University Institutional Review Board. The approval number is QU-IRB 1174-EA/19. The evaluation experiments were carried out on the ground floor of the 'B09' building of Qatar University. Figure 10 illustrates the floor plan of the building 'B09' (ground floor).

We selected 10 people including 8 females and 2 males to evaluate the navigation systems in real-time. The age of participants is ranged from 22 to 39 (Mean = 29.30, standard deviation (SD) = 4.90). Since the participants were sighted, during experiments we made them wear blindfolds.

The blindfolded participants were asked to walk from the entrance door of the B09 building to two specific points of interest in the B09 building. Each participant has to walk from point A to B (Red line in the floor plan, distance = 30 m) and A to C (Blue line in the floor plan, distance = 47 m) using the three navigation systems separately. We recorded all the walking experiments using a video camera for further assessments.

To evaluate the performance of the systems and compare their merits and demerits, we mainly focused on computing the time required for the navigation using each system, error committed by the user in each navigation trail while using different systems and users' feedback about the three systems.

Navigation analysis

In this section, a study was conducted to examine the performance of the navigation systems in terms of time taken for traveling and errors that occurred during navigation. Ten blindfolded participants including 8

females and 2 males were asked to walk from point A to B and point A to C (shown in Fig. 10) using three navigation systems separately. All together each participant has to do six navigation experiments which include navigation through two routes (A to B and A to C) by using each of the three navigation systems separately.

For CamNav and QRNav systems, a Samsung Galaxy S7 smartphone was employed for running the android application. The Wi-Fi network of Qatar university was utilized as a medium for sending the images from the Android application to the server. On the server-side, an Asus ROG edition laptop with 24 GB ram and Nvidia GTX 1060 GPU was utilized for processing the captured images. BLE beacon-based system was directly implemented in the Samsung Galaxy S7 smartphone. Eleven BLE beacons are used to cover the whole navigation route. The positioning and hardware specifications of the beacons are detailed in the system overview section. Figure 8 shows the distribution of beacons over the experimental area.

Two metrics; time taken for navigation and error occurred during navigation in terms of the number of steps are considered in this navigation analysis to compare the performance of CamNav, QRNav, and BLE based system. The procedure followed for conducting the experiment is detailed as follows. Eyes of each participant are covered with a cloth prior to the navigation experiment. The details, as well as name of the source and destination of navigation routes, are not revealed to participants since they are familiar with the indoor areas of the building where the experiment is conducted. A participant is randomly called from 10 selected participants and asked to do the experiment in an order such that, walk from point A to B and A to C in two different trails using BLE application initially. Then we made the same participant repeat the experiment for two routes using the CamNav system followed by the QRNav system. The same procedure was repeated for the remaining nine participants. During each navigation trail, the time taken for navigation was measured using a stopwatch timer. Moreover, the error occurred during the navigation trail was measured manually in terms of the number of steps. In order to measure the error occurred during navigation, we considered some predefined points in the navigation route as location reference points. We considered four reference points in route 1 and seven reference points in route 2. The error occurred during navigation is computed as the sum of the number of steps the user is behind or ahead with respect to each predefined location reference points in the navigation route during the navigation by using each navigation system. The average time (in seconds) required by each navigation system for each route was displayed in Table 4. Route 1 represents point A to B and route 2 represents A to C.

The average navigation time taken by participants while using BLE based system was comparatively less compared to the other two systems, CamNav and QRNav. The average time taken by the BLE based system for the shorter route (route A to B) and longer route (route A to C) was 106.8 s with a standard deviation of 16.88 and 174.3 s with a standard deviation of 23.19 respectively. While comparing with BLE based system, the CamNav system took 17 s and 23 s more average time in routes 1 and 2 respectively. The average time taken by the QRNav system is almost similar to the CamNav system. The main reason for the difference in average time while using different systems are, the client-server communication used in CamNav and QRNav consumes time, but less than 2 s for processing each image. Moreover, CamNav and QRNav provide abundant information regarding the current location and surroundings such as information about doors or walls. So the user has to wait for some seconds to hear and understand all the instructions. But the BLE beacon-based system will provide only the limited instructions to reach the destination. While using the QRNav and CamNav systems, we have noted that the results are affected by the walking speed of the user. This is because when the user is walking at a higher speed, the chance of getting blurry pictures is higher. In fact, the average speed of people with visual impairments varies according to their prior knowledge of the area they are walking in. BLE beacons based system enabled the user to walk a little bit fast and thereby reduced the navigation time. BLE beacons based system is able to provide real-time navigation service, while QRNav and CamNav are experiencing a small delay up to 2 s.

The average error obtained in each system is illustrated in Table 5.

The average errors in terms of the number of steps were less in CamNav and QRNav compared to BLE based system. The average error occurred in CamNav was 3.1 with a standard deviation of 0.56 and 6.1 with a standard deviation of 1.10 respectively for routes 1 and 2. In the QRNav system similar average value was obtained, 3.3 with a standard deviation of 0.48 in route

Table 4 Average navigation time (standard deviation in brackets)

Systems	Average navigation time in seconds	
	Route 1	Route 2
CamNav	123.2 (11.66)	197.6 (18.31)
QRNav	117.2 (19.22)	204.1(16.33)
BLE APP	106.8 (16.88)	174.3 (23.19)

Table 5 Average error in terms of the number of steps (Standard deviation in brackets)

Systems	Average error	
	Route 1	Route 2
CamNav	3.1 (0.56)	6.1 (1.10)
QRNav	3.3 (0.48)	5.5 (0.84)
BLE APP	4.3 (0.94)	8.7 (1.33)

Table 6 Navigation efficiency index recorded in the systems

System	NEI
CamNav	1.010
QRNav	0.996
BLE APP	1.019

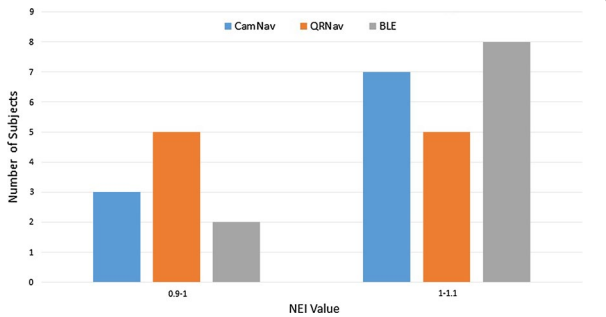


Fig. 11 Distribution of navigation efficiency index

1 and 5.5 with a standard deviation of 0.84 in route 2. It proved that CamNav and QRNav are more accurate in providing reliable navigation services for users. The accuracy of the BLE beacon-based system was around 2-3 meters that raised conflicts about the user's position while moving through the junctions. This conflict in position estimation generated uncertainty in provided navigation instructions and thereby misguided the user.

Navigation efficiency index (NEI) [62] was included to evaluate the navigation performance of the systems. NEI is defined as the ratio of actual traveled path's distance to optimal path's distance between source and destination. The NEI obtained in each system are displayed in Table 6. The obtained NEI in each system are acceptable and it displayed the effectiveness of navigation instructions provided by each system.

Distribution of NEI values among participants of navigation experiment for each system are shown in Fig. 11.

Table 7 Rating of functionalities in the systems

Functionalities	Navigation systems		
	CamNav	QRNav	BLE APP
Navigation path suggestion	4.3	4.5	3.5
Location estimation	4.3	4.8	3.8
Speech instructions	4	4	4

Qualitative analysis

In order to examine the effectiveness of three systems, we prepared a questionnaire for the people who participated in the experiments. After completing the real-time experiments, we asked the participants to fill out the questionnaire. The questionnaire contains two sections. In the initial section, the participants are asked to rate the main functionalities (Navigation path suggestion, Location estimation, and Speech instructions) of the navigation systems using a Likert scale of points 1 (very bad) to 5 (excellent). The average rating of the functionalities is shown in Table 7.

"Related work" section of the questionnaire is dedicated to evaluating the satisfaction of the participants, the effectiveness of three navigation systems. In order to evaluate the satisfaction of the participants and effectiveness, we adopted a popular method known as the System Usability Scale (SUS) which has been commonly used in the literature for the usability evaluation task for various human-computer interaction systems. SUS is a simple, "quick and dirty" reliable tool to evaluate the usability of the systems. The high reliability, as well as the validity of the SUS tool, made it popular among the human-computer interaction community. The SUS tool consists of ten questions. Out of the ten questions, five (questions having odd serial numbers) are positive items and the remaining five (questions having even serial numbers) are negative items. The questions of the SUS questionnaire tool are displayed in Table 8.

The participants are asked to respond and record their agreement for each question with a 5 point Likert scale where 1 means strongly disagree and 5 means strongly agree to the respective statements. Once the participants record their rating in terms of values, following equation is utilized to compute the overall SUS score of the system.

$$SUS\ score = 2.5 \times \left[\sum_{n=1}^{5}(R_{2n-1} - 1) + (5 - R_{2n})\right] \quad (5)$$

In Eq. (1) Ri means the rating value (1 to 5) given for the ith question by the participant and n is the serial number of question. Overall SUS score calculation follows a

Table 8 SUS questionnaire tool

I think that I would like to use this system frequently
Found the system unnecessarily complex
I thought the system was easy to use
I think that I would need the support of a technical person to be able to use this system
I found the various functions in this system were well integrated
I thought there was too much inconsistency in this system
I would imagine that most people would learn to use this system very quickly
I found the system very cumbersome to use
I felt very confident using the system
I needed to learn a lot of things before I could get going with this system

Table 9 System usability scores

	Navigation systems		
	CamNav	QRNav	BLE APP
Average overall SUS score	84.3	88	76.75
Standard deviation of SUS score	7.79	5.98	7.64
Cronbach's alpha	0.73	0.61	0.61

simple procedure. For each odd number question or positive item, 1 will be subtracted from the rating value given by the participant. For each even number questions or negative item, the rating value given by the participants is subtracted from a constant value 5. Once the above mentioned subtraction operations are done, the obtained final values for each question are added and sum of these values are multiplied by 2.5. The overall SUS score ranges between 0 to 100, where higher value indicate superior performance. The average SUS score obtained for each system , standard deviation and Cronbach's alpha of the SUS questionnaire for each system are displayed in Table 9. The Cronbach's alpha is a commonly used statistical reliability analysis method to measure the internal consistency of questionnaire.

The average usability score based on the SUS questionnaire for the CamNav system is 84.5 with a standard deviation of 7.79. QRNav obtained an average score of 88 with a standard deviation of 5.98 and which is the highest among all the three systems. BLE based application obtained the least average score (76.75 with a standard deviation of 7.64). While comparing these three systems based on the usability score, QRNav and CamNav displayed superior performance. Despite the fact that the BLE app obtained the least score, the SUS score obtained for each system including the BLE app are acceptable since the values are above 70. The Cronbach alpha which indicates the reliability of scores obtained in the questionnaire is 0.73 for SUS questionnaire in the CamNav system. Incase of QRNav and BLE based application the Cronbach alpha is 0.61. The obtained Cronbach alpha value shows the strong reliability for the SUS tool used in usability evaluation.

Discussion

The results obtained from the real-time experiments and feedback from the participants show the effectiveness of the CamNav and QRNav systems over the BLE beacon-based navigation system. All the participants reached their destination with minimal external assistance. Some of the participants relied on external assistance while passing through the junctions in the navigation route. The external assistance was required in the junctions, when the user is turning to his right or left side. Instead of turning 90°, sometimes, they are turning more than that and it will lead to a condition where the user is deviating from the original path. While using BLE based application, it has noted that the is user bumped to the wall or door sometimes. But this issue was not present in the QRNav and CamNav systems. Because the QRNav is able to detect the doors or walls with the help of QR codes attached to it. CamNav is trained to recognize the walls and door areas in the indoor region.

It is clear from the user's feedback that the participants found that Android application was easy to use. Minimal training is only required for using the Android application. Any user with minimum knowledge of smartphone can use this application for navigation. In case of QRNav and CamNav user has to keep the mobile straight or down for location recognition. Some user has found that it is difficult to position the smartphone in a fixed direction for long time. The majority of the participants said that they will suggest the CamNav or QRNav system for their visually impaired friend or relatives. However the current version of CamNav and QRNav provides information about the doors in the navigation route, the participants suggested adding a module to detect and recognize the static (e.g. table, chair, etc.) and dynamic (e.g. people) objects or obstacles to ensure safe navigation.

The participants for the field test has been randomly selected and the criteria for selecting the participants was minimum knowledge of smartphone and English language understanding. The selected participants are not having any physical disability as well as hearing disability. since they have to walk properly by closing eyes with the help of voice instructions. To best of our knowledge , the participant selection has not adversely effected the result. During experiments a minimal help was provide for blindfolded users in order to preserve their safety. For each participants, all six navigation trails was conducted

Table 10 Comparison of developed indoor navigation systems

Reference	Technology	System	Techniques	Tested by	Test	Feedback	Accuracy	Remarks
CamNav	Computer vision	Smartphone, laptop	CNN based scene recognition	Blindfolded	Error occurred during navigation and navigation time	Voice	Medium	(—) Hard to differentiate indoor locations with similar appearance
QRNav	Computer vision	Smartphone, laptop, QR codes	QR code recognition	Blindfolded	Error occurred during navigation and navigation time	Voice	Medium	(—) Requires large amount of QR codes for safe navigation
BLE base system	BLE	Beacons and smartphone	Fingerprinting and trilateration	Blindfolded	Error occurred during navigation and navigation time	Voice	Low	(—) Relatively low accuracy resulted in navigation errors

on a same day, with providing proper rest between each trail. All of the three system are tested by same group of 10 participants. The device and instruments used for experiments were same for all participants during each trail. Moreover same tool (navigation analysis and usability evaluation using SUS tool) was used for assessing the performance of three systems. Despite the fact that the participants are not visually impaired, we believe the obtained result are enough to compare the three different approaches for indoor navigation.

While comparing three different approaches we found that the QR code recognition is very quick and precise. It doesn't require powerful processors or devices to carry out the QR code decoding process. On other hand CNN based scene recognition is effective for identifying the doors or walls or other static obstacle in the navigation path. Same can be achieved in QR code system, but it requires pasting the QR codes in walls and doors. And it is not practically possible to paste QR codes in all walls and doors since it is not aesthetically pleasing for other people with normal vision. On the other hand CNN recognition may fail in locations with similar appearance. The BLE beacons are not much accurate to estimate the exact position of the user. But it can be used to identify the rooms or corridor where the user is present. While navigating through straight paths in corridors BLE beacons are effective for getting real time position of the user. All of the three approaches have merits as well as limitations. But three approaches can be integrated to develop a hybrid navigation system for people with visual impairments where merit of one system will compensate the limitation of other such as QR code recognition or BLE beacons positioning can solve the issue of recognizing locations with similar appearance using CNN model.

We already provided the comparison of some existing indoor navigation solutions for people with VI in Table 2. With respect to Table 2 , here we provide the comparison of three developed systems in Table 10 while considering the same attributes used in Table 2.

Conclusion

In this article, we developed three indoor navigation systems for people with visual impairments. The proposed systems utilize different technologies, CNN based scene recognition, QR code recognition and BLE based positioning to navigate people with visual impairments. The three systems has been implemented and assessed in indoor environment. BLE based system took less time for navigation, while QRNav and CamNav showed better performance while considering the average error obtained in each system and system usability scores.

Three systems was tested on blindfolded people in a real-time environment. The average error obtained in CamNav is 3.1 steps and 6.1 steps with a standard deviation of 0.56 and 1.10 respectively in route 1 and route 2. The average error was highest in BLE based application with a value 4.3 steps and 8.7 steps with standard deviation 0.94 and 1.33 in routes 1 and 2 respectively. In usability experiments , QRNav obtained highest system usability score of 88, CamNav got a score of 84.3 and BLE based system got an average score of 76.75. Indoor scene recognition and QR code recognition are affected by the walking speed of the user. Embedding a obstacle detection sensor along with navigation system can ensure the safety of user during the presence of mobile obstacle. In the future, we will consider integrating QR code recognition in CamNav system to cope with the identification of location which is similar in appearance. The future version of CamNav will address most of the requirements suggested by the users who participated in the real-time evaluation.

Acknowledgements
Not applicable.

Authors' contributions
JK and AK mainly contributed in the development of navigation systems. JK was a major contributor in writing the manuscript. The concept of the work was contributed by SA. SA was the principal investigator of the project. SA as well as AK revised and edited the manuscript.AA provided guidance and assistance to improve the usability of system since psychology has a substantial role in the development of assistive systems. All authors read and approved the final manuscript.

Funding
This publication was supported by Qatar University Collaborative High Impact Grant QUHI-CENG-18/19-1.

Availability of data and materials
Up on reasonable request, the corresponding author will make the datasets available.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Ethics approval and consent to participate
This study was approved by the Qatar University Institutional Review Board. The Approval Number is QU-IRB 1174-EA/19.

Author details
[1] Department of Computer Science and Engineering, Qatar University, Al Jamiaa Street, Doha, Qatar. [2] Department of Psychological Sciences, Qatar University, Al Jamiaa Street, Doha, Qatar.

Received: 17 August 2019 Accepted: 30 November 2019
Published online: 11 December 2019

References

1. Fallah N, Apostolopoulos I, Bekris K, Folmer E. Indoor human navigation systems: a survey. Interact Comput. 2013;25(1):21–33.
2. Maghdid HS, Lami IA, Ghafoor KZ, Lloret J. Seamless outdoors-indoors localization solutions on smartphones: implementation and challenges. ACM Comput Surv (CSUR). 2016;48(4):53.
3. Koyuncu H, Yang SH. A survey of indoor positioning and object locating systems. IJCSNS Int J Comput Sci Netw Secur. 2010;10(5):121–8.
4. Fangmeyer JR JA, Rosales CV, Rodríguez DM, Pinero RFB. Evolution of indoor positioning technologies: a survey. 2018.
5. Ben-Afia A, Deambrogio L, Salós D, Escher AC, Macabiau C, Soulier L, et al. Review and classification of vision-based localisation techniques in unknown environments. IET Radar Sonar Navig. 2014;8(9):1059–72.
6. Jeon KE, She J, Soonsawad P, Ng PC. BLE beacons for Internet of Things applications: survey, challenges, and opportunities. IEEE Internet Things J. 2018;5(2):811–28.
7. Chang K. Bluetooth: a viable solution for IoT? [Industry Perspectives]. IEEE Wireless Commun. 2014;21(6):6–7.
8. Yin H, Jiao X, Chai Y, Fang B. Scene classification based on single-layer SAE and SVM. Expert Syst Appl. 2015;42(7):3368–80.
9. Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: A 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell. 2017;40(6):1452–64.
10. Hightower J, Borriello G. Location systems for ubiquitous computing. Computer. 2001;8:57–66.
11. Iglesias HJP, Barral V, Escudero CJ. Indoor person localization system through RSSI Bluetooth fingerprinting. In: 2012 19th international conference on systems, signals and image processing (IWSSIP). New York: IEEE; 2012; p. 40–43.
12. Clark BK, Winkler EA, Brakenridge CL, Trost SG, Healy GN. Using Bluetooth proximity sensing to determine where office workers spend time at work. PloS ONE. 2018;13(3):e0193971.
13. Zandbergen PA, Barbeau SJ. Positional accuracy of assisted GPS data from high-sensitivity GPS-enabled mobile phones. J Navig. 2011;64(3):381–99.
14. Lindeberg T. Scale invariant feature transform. Scholarpedia. 2012;7(5):10491.
15. Speeded-Up Robust Features (SURF). Computer vision and image understanding. 2008;110(3):346 – 359. Similarity Matching in Computer Vision and Multimedia. http://www.sciencedirect.com/science/article/pii/S1077314207001555.
16. Wang H, Zhang S. Evaluation of global descriptors for large scale image retrieval. In: International conference on image analysis and processing. Berlin: Springer; 2011. p. 626–635.
17. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436.
18. Lee YH, Medioni G. Wearable RGBD indoor navigation system for the blind. In: Agapito L, Bronstein MM, Rother C, editors. Computer vision—ECCV 2014 workshops. Cham: Springer International Publishing; 2015. p. 493–508.
19. Kamisaka D, Muramatsu S, Iwamoto T, Yokoyama H. Design and implementation of pedestrian dead reckoning system on a mobile phone. IEICE Trans Inf Syst. 2011;94(6):1137–46.
20. Ban R, Kaji K, Hiroi K, Kawaguchi N. Indoor positioning method integrating pedestrian Dead Reckoning with magnetic field and WiFi fingerprints. In: 2015 Eighth international conference on mobile computing and ubiquitous networking (ICMU); 2015. p. 167–72.
21. Woodman OJ. An introduction to inertial navigation. University of Cambridge, Computer Laboratory; 2007. UCAM-CL-TR-696. https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-696.pdf.
22. Bouet M, dos Santos AL. RFID tags: Positioning principles and localization techniques. In: 2008 1st IFIP wireless days; 2008; p. 1–5.
23. Fu Q, Retscher G. Active RFID trilateration and location fingerprinting based on RSSI for pedestrian navigation. J Navig. 2009;62(2):323–40.
24. He S, Chan SG. Wi-Fi fingerprint-based indoor positioning: recent advances and comparisons. IEEE Commun Surv Tutor. 2016;18(1):466–90.
25. Farid Z, Nordin R, Ismail M. Recent advances in wireless indoor localization techniques and system. J Comput Netw Commun. 2013;2013:12.
26. Orujov F, Maskeliūnas R, Damaševičius R, Wei W, Li Y. Smartphone based intelligent indoor positioning using fuzzy logic. Future Gener Comput Syst. 2018;89:335–48.
27. Do TH, Yoo M. An in-depth survey of visible light communication based positioning systems. Sensors. 2016;16(5). http://www.mdpi.com/1424-8220/16/5/678.
28. Alarifi A, Al-Salman A, Alsaleh M, Alnafessah A, Al-Hadhrami S, Al-Ammar MA, et al. Ultra wideband indoor positioning technologies: analysis and recent advances. Sensors. 2016;16(5). http://www.mdpi.com/1424-8220/16/5/707.
29. Tian Y, Yang X, Yi C, Arditi A. Toward a computer vision-based wayfinding aid for blind persons to access unfamiliar indoor environments. Mach Vis Appl. 2013;24(3):521–35.
30. Deniz O, Paton J, Salido J, Bueno G, Ramanan J. A vision-based localization algorithm for an indoor navigation app. In: 2014 Eighth international conference on next generation mobile apps, services and technologies. New York: IEEE; 2014. p. 7–12.
31. Canny J. A computational approach to edge detection. In: Readings in computer vision. Amsterdam: Elsevier; 1987. p. 184–203.
32. Garcia G, Nahapetian A. Wearable computing for image-based indoor navigation of the visually impaired. In: Proceedings of the conference on Wireless Health. New York: ACM; 2015. p. 17.
33. Manlises C, Yumang A, Marcelo M, Adriano A, Reyes J. Indoor navigation system based on computer vision using CAMShift and D* algorithm for visually impaired. In: 2016 6th IEEE international conference on control system, computing and engineering (ICCSCE). New York: IEEE; 2016; p. 481–4.
34. Tian Y, Yang X, Arditi A. Computer vision-based door detection for accessibility of unfamiliar environments to blind persons. In: International conference on computers for handicapped persons. Berlin: Springer; 2010. p. 263–70.
35. Costa P, Fernandes H, Martins P, Barroso J, Hadjileontiadis LJ. Obstacle detection using stereo imaging to assist the navigation of visually impaired people. Procedia Comput Sci. 2012;14:83–93.
36. Lee YH, Medioni G. RGB-D camera based wearable navigation system for the visually impaired. Computer Vis Image Understand. 2016;149:3–20.
37. Li B, Muñoz JP, Rong X, Chen Q, Xiao J, Tian Y, et al. Vision-based mobile indoor assistive navigation aid for blind people. IEEE Trans Mob Comput. 2018;18(3):702–14.

38. Kanwal N, Bostanci E, Currie K, Clark AF. A navigation system for the visually impaired: a fusion of vision and depth sensor. Appl Bionics Biomech. 2015; 2015.
39. Idrees A, Iqbal Z, Ishfaq M. An efficient indoor navigation technique to find optimal route for blinds using QR codes. In: 2015 IEEE 10th conference on industrial electronics and applications (ICIEA). New York: IEEE; 2015; p. 690–5.
40. Al-Khalifa S, Al-Razgan M. Ebsar: indoor guidance for the visually impaired. Comput Electr Eng. 2016;54:26–39.
41. Gude R, Østerby M, Soltveit S. Blind navigation and object recognition. Denmark: Laboratory for Computational Stochastics, University of Aarhus; 2013.
42. Legge GE, Beckmann PJ, Tjan BS, Havey G, Kramer K, Rolkosky D, et al. Indoor navigation by people with visual impairment using a digital sign system. PloS ONE. 2013;8(10):e76783.
43. Zeb A, Ullah S, Rabbi I. Indoor vision-based auditory assistance for blind people in semi controlled environments. In: 2014 4th international conference on image processing theory, tools and applications (IPTA). New York: IEEE; 2014. p. 1–6.
44. Huang YC, Ruan SJ, Christen O, Naroska E. A wearable indoor locating system based on visual marker recognition for people with visual impairment. In: 2016 IEEE 5th global conference on consumer electronics. New York: IEEE; 2016. p. 1–2.
45. Raj CR, Tolety S, Immaculate C. QR code based navigation system for closed building using smart phones. In: 2013 International mutliconference on automation, computing, communication, control and compressed sensing (iMac4s). New York: IEEE; 2013. p. 641–4.
46. Chuang CH, Hsieh JW, Fan KC. A smart handheld device navigation system based on detecting visual code. In: 2013 international conference on machine learning and cybernetics. vol. 3. New York: IEEE; 2013; p. 1407–12.
47. Rituerto A, Fusco G, Coughlan JM. Towards a sign-based indoor navigation system for people with visual impairments. In: Proceedings of the 18th international ACM SIGACCESS conference on computers and accessibility. New York: ACM; 2016. p. 287–8.
48. Ahmetovic D, Gleason C, Ruan C, Kitani K, Takagi H, Asakawa C. NavCog: a navigational cognitive assistant for the blind. In: Proceedings of the 18th international conference on human-computer interaction with mobile devices and services. New York: ACM; 2016. p. 90–9.
49. Kim JE, Bessho M, Kobayashi S, Koshizuka N, Sakamura K. Navigating visually impaired travelers in a large train station using smartphone and Bluetooth low energy. In: Proceedings of the 31st annual ACM symposium on applied computing. New York: ACM; 2016. p. 604–11.
50. Cheraghi SA, Namboodiri V, Walker L. GuideBeacon: Beacon-based indoor wayfinding for the blind, visually impaired, and disoriented. In: 2017 IEEE international conference on pervasive computing and communications (PerCom). New York: IEEE; 2017. p. 121–30.
51. Bilgi S, Ozturk O, Gulnerman AG. Navigation system for blind, hearing and visually impaired people in ITU Ayazaga campus. In: 2017 international conference on computing networking and informatics (ICCNI). New York: IEEE; 2017. p. 1–5.
52. Duarte K, Cecílio J, Furtado P. Easily guiding of blind: Providing information and navigation-SmartNav. In: International wireless internet conference. Berlin: Springer; 2014. p. 129–34.
53. Murata M, Ahmetovic D, Sato D, Takagi H, Kitani KM, Asakawa C. Smartphone-based localization for blind navigation in building-scale indoor environments. Pervas Mob Comput. 2019;57:14–32.
54. AL-Madani B, Orujov F, Maskeliūnas R, Damaševičius R, Venčkauskas A. Fuzzy logic type-2 based wireless indoor localization system for navigation of visually impaired people in buildings. Sensors. 2019;19(9):2114.
55. Snyder B, Bosanac D, Davies R. Introduction to apache activemq. Active MQ in action. 2017. p. 6–16.
56. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467. 2016.
57. CamNav-dataset. https://github.com/akarkar/CamNav-dataset. Accessed 30 Apr 2019.
58. Mapbox. https://www.mapbox.com. Accessed 6 Nov 2019.
59. Zxing. http://code.google.com/p/zxing/. Accessed 4 Nov 2019.
60. Zbar bar code reader. http://zbar.sourceforge.net/. Accessed 25 Apr 2019.
61. Shchekotov M. Indoor localization method based on Wi-Fi trilateration technique. In: Proceeding of the 16th conference of fruct association; 2014. p. 177–9.
62. Ganz A, Schafer J, Gandhi S, Puleo E, Wilson C, Robertson M. PERCEPT indoor navigation system for the blind and visually impaired: architecture and experimentation. Int J Telemed Appl. 2012;2012:19.

Publisher's Note

sensors

MDPI

Article

Automatic Object Detection Algorithm-Based Braille Image Generation System for the Recognition of Real-Life Obstacles for Visually Impaired People

Dayeon Lee and Jinsoo Cho *

IT Convergence Engineering and Computer Convergence Major, Gachon University, Seongnam 13120, Korea; lidy030@gachon.ac.kr

* Correspondence: jscho@gachon.ac.kr

Abstract: The global prevalence of visual impairment due to diseases and accidents continues to increase. Visually impaired individuals rely on their auditory and tactile senses to recognize surrounding objects. However, accessible public facilities such as tactile pavements and tactile signs are installed only in limited areas globally, and visually impaired individuals use assistive devices such as canes or guide dogs, which have limitations. In particular, the visually impaired are not equipped to face unexpected situations by themselves while walking. Therefore, these situations are becoming a great threat to the safety of the visually impaired. To solve this problem, this study proposes a living assistance system, which integrates object recognition, object extraction, outline generation, and braille conversion algorithms, that is applicable both indoors and outdoors. The smart glasses guide objects in real photos, and the user can detect the shape of the object through a braille pad. Moreover, we built a database containing 100 objects on the basis of a survey to select objects frequently used by visually impaired people in real life to construct the system. A performance evaluation, consisting of accuracy and usefulness evaluations, was conducted to assess the system. The former involved comparing the tactile image generated on the basis of braille data with the expected tactile image, while the latter confirmed the object extraction accuracy and conversion rate on the basis of the images of real-life situations. As a result, the living assistance system proposed in this study was found to be efficient and useful with an average accuracy of 85% a detection accuracy of 90% and higher, and an average braille conversion time of 6.6 s. Ten visually impaired individuals used the assistance system and were satisfied with its performance. Participants preferred tactile graphics that contained only the outline of the objects, over tactile graphics containing the full texture details.

Keywords: image processing; object detection; artificial intelligence; blind; braille system

Citation: Lee, D.; Cho, J. Automatic Object Detection Algorithm-Based Braille Image Generation System for the Recognition of Real-Life Obstacles for Visually Impaired People. *Sensors* **2022**, *22*, 1601. https://doi.org/10.3390/s22041601

Academic Editors: Hossein Anisi, Vahid Abolghasemi and Saideh Ferdowsi

Received: 29 December 2021
Accepted: 25 January 2022
Published: 18 February 2022

1. Introduction

The leading cause of visual impairment can be congenital or a result of accidents, aging, or diseases. In addition, the number of people with acquired vision loss is increasing because of urban environmental factors resulting from the development of electronic devices [1,2]. A survey made by the World Health Organization (WHO) in 2020 indicated that approximately 2.2 billion people, which accounts for 28.22% of the global population, are visually impaired (i.e., near or distance visual impairment) [3,4].

Visually impaired people rely on their auditory perception and somatosensation—primarily sound and braille—to obtain information from the environment; they use assistive devices such as canes to recognize obstacles. However, although 28.22% of the global population accounts for visually impaired individuals [5,6], accessible facilities are not universally installed, leading to issues of social discrimination due to the limitations of their activities. Particularly, they cannot face unexpected situations outdoors independently, thereby restricting their activities to indoors or in their neighborhood. Accessible facilities such

as tactile pavements and tactile signs are not appropriately installed in all institutions. Moreover, some countries do not provide support for assistive devices. In addition, most artworks, such as paintings and sculptures, cannot be touched to preserve them, making it difficult for visually impaired people to enjoy cultural activities through their imagination alone with tactile brochures. Therefore, researchers conducted numerous studies to help them become self-sufficient in their daily lives. In particular, studies on providing information via braille have recently gained attention. However, most of these studies focused on tactile maps or graphic image braille conversion. A system is needed worldwide to ease their daily lives because it is difficult to assist the visually impaired individuals in real life.

This study proposes a living assistance system based on images of the surroundings and objects that visually impaired people want to experience in real life that are captured by smart glasses. The system stores object information using an object detection algorithm to provide voice guidance when the user goes outdoors. Moreover, the system provides an object image braille conversion service using an object extraction algorithm when indoors and carrying a braille pad. The braille data are generated as binary data to enable use in various braille pads, and the images are generated at three degrees of expression to enable users to recognize the shapes at different types. The accuracy of the proposed system is calculated by comparing the example tactile image with the expected tactile image on the basis of the braille data, and the usefulness of the system is evaluated by comparing the object detection results in real-life images and the execution time.

2. Related Research

Researchers conducted various studies regarding the living assistance for visually impaired people. Previous studies were focused on the generation of tactile signs and maps as navigation aids for the visually impaired, image conversion, and the development of tactile image output devices for braille pads. However, there is a lack of studies on the generation of tactile images based on real-life images or systems that assist with real-life outdoor activities, such as the automatic object detection voice guidance system proposed in this study.

2.1. Similar Research

2.1.1. Tactile Graphics

Tactile maps and images are generated through image processing based on general maps to create tactile maps. Tactile maps are the most provided navigation aid for the visually impaired people by public institutions. However, tactile maps are gradually being provided by various institutions, fueling further research on their development.

Kostopoulos et al. [7] proposed a method for generating tactile maps based on a map image created by reading the road names written on a map via OCR and converting it into a road image, as shown in Figure 1. Although the proposed system for creating tactile maps can quickly recognize roads on the basis of the road names, it cannot detect alleys without a name. Moreover, OCR is slow and limited although it is faster than the existing algorithms.

Zeng et al. [8] developed an interactive map in which the user can zoom in and out, as shown in Figure 2. They allowed users to explore the tactile map by dividing it into zoom levels. However, a post-experiment survey found that visually impaired people preferred maps with only two zoom levels, and the usage time increased due to various factors such as the production of the interactive map, the guidance of the selections, and the selection.

Moreover, Krufkaf et al. [9] proposed an advanced braille conversion algorithm for vector graphics on the basis of previous studies. The algorithm extracted object boundaries using the outline information of the graphic based on the vector graphics hierarchical characteristics. The levels are classified on the basis of the extracted boundaries, and the multi-level braille is converted to a braille tablet using the tiger advantage braille printer program [10]. Although the proposed multi-level braille conversion system can provide meaningful results, it is difficult to apply to real-life objects using vector graphics, as shown in Figure 3.

(a) Primary map image (b) Obtained road network structure (c) Obtained 3D map model

Figure 1. Map image-based tactile map production method [7].

(a) The screenshots of Map2 on iPad

(b) Screenshots of Map3 on the HBMap system

Figure 2. iPad and HBMap system-based interactive maps [8].

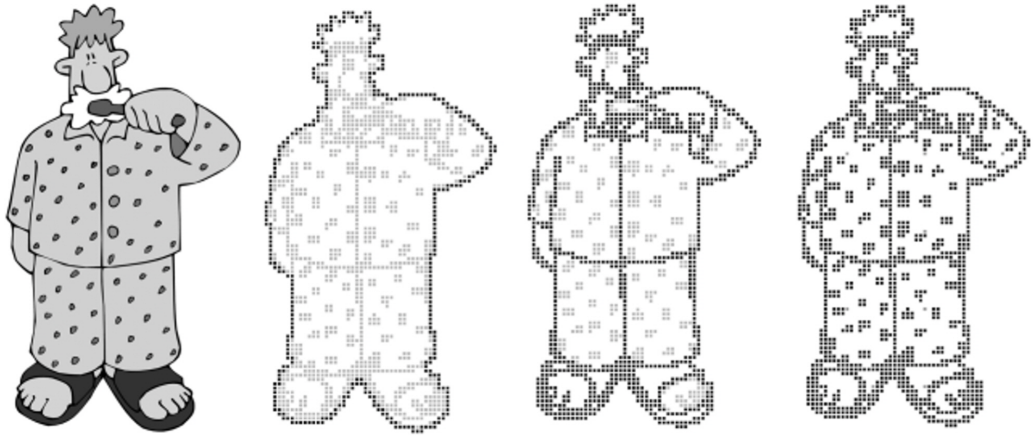
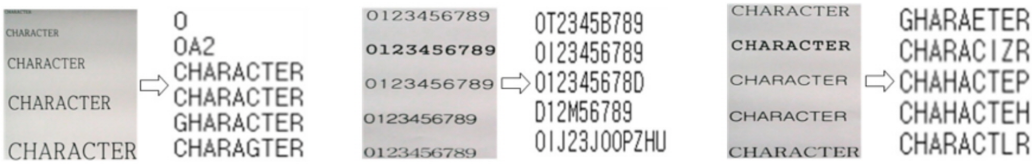


Figure 3. Outputs of proposed method for the vector graphic [9].

In Korea, Kim et al. [11] investigated braille conversion on the basis of images captured via a webcam. The locations with and without data are compared to identify characters

in the image by analyzing the images using MATLAB. Figure 4 shows the evaluation of the recognition level according to the font size, font type, and camera performance. In addition, an algorithm was developed by configuring an optimal environment based on the evaluation results. Although their research showed significant results, the system can only convert numbers and uppercase English letters, and it cannot identify objects other than letters or recognize Korean letters.



(a) Alphanumeric converison by size (b) Numeric converison by font (c) Alphanumeric conversion by font

Figure 4. Image conversion according to font size and font [11].

Lee et al. [12] developed a banknote recognition system using Raspberry Pi as a camera. The process consisted of two steps (i.e., extraction and matching). The researchers compared the extraction algorithms SIFT, SURF, and ORB; they adopted SIFT because it yielded the highest recognition rate. The system achieved high accuracy even when changing the shooting method or in unsuitable environments (e.g., low light or rotated banknote) by generating vector images using extreme values as features. Nevertheless, the brute-force algorithm requires extensive time for recognition, as shown in Figure 5, making it unsuitable for this study, which uses many objects.

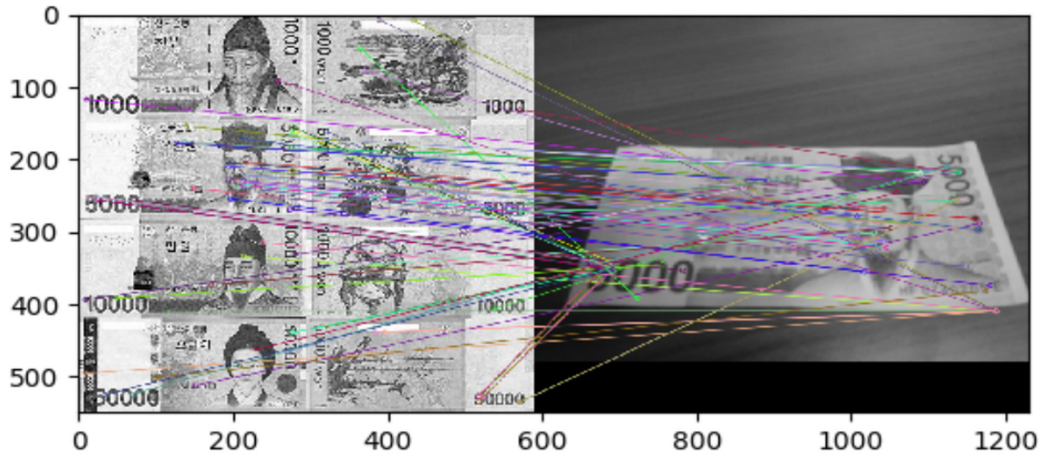


Figure 5. Keypoints Matching Using the Brute-Force Algorithm [12].

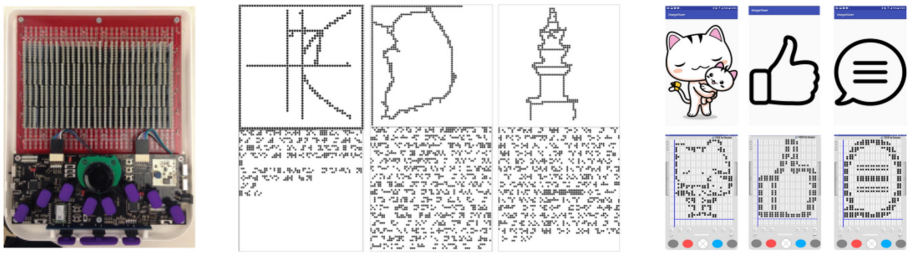
2.1.2. Braille Pad

Researchers have made several attempts to output tactile images by combining a haptic device with a braille display [13].

Kim, S. et al. [14,15] proposed a 2D braille display to output data in the digital accessible information system (DAISY) and the electronic publication (EPUB) formats. They developed the braille pad for outputting braille information and the technology for tactile image conversion, as shown in Figure 6. Tactile image tests were conducted using simulators, and the tactile image conversion technology quantizes and binarizes data to convert graphs, graphic images, and even photos, enabling them to obtain significant results.

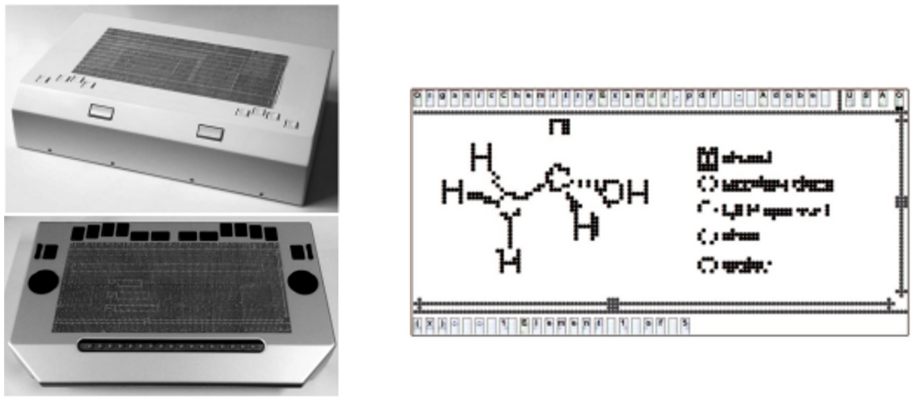
Prescher et al. [16] proposed a PDF-editor-based braille pad and braille conversion system. The user interface (UI) for displaying and editing PDF content was designed to show on one screen using a horizontally long touch-enabled braille pad. As both the content and editing UI are displayed on one screen, excessive information is provided at once, making it difficult for first-time users. Moreover, it can only translate the diagrams

and text input, which are in PDF files rather than images, although it can display diagrams as shown in Figure 7.



(a) braille pad prototype (b) Braille pad simulator output screen (c) braille pad simulator

Figure 6. Braille pad prototype and Output screen [14,15].



(a) Braille pad example (b) Output example

Figure 7. Braille pad and Output example [16].

2.1.3. Supplementation and Service

In addition, various products and services are being researched to assist the visually impaired.

Kłopotowska et al. [17] studied architectural typhlographics and developed them through multi-criteria analysis by integrating the characteristics of braille maps and architectures (Figure 8). The study results show the future growth potential of typhlographics on the basis of its social values of enabling tourism for the visually impaired in addition to its broad utility in the development of tactile architectural drawings such as diversification of architectural education and interior design.

Morad [18] studied the assistive devices that receive location coordinates via the global positioning system (GPS) and process data through a PIC controller to output specific voice messages stored in the device for visually impaired people. The study aimed to develop an affordable and easy-to-use assistive device that helps the visually impaired people find their way on their own as they listen to the voice messages through the headset. It received a positive response from them when the device was used by people with visual impairments.

On the other hand, Fernandes et al. [19] proposed a radiofrequency identification (RFID)-based cane navigation system to guide people with visual impairments by using the RFID device installed under the road. The navigation system provides audio navigation assistance to reach the desired destination through the route calculation and location tracking using the RFID tags once the user inputs a specific destination in the cane. It is considered to have a significant growth potential owing to its higher accuracy than GPS and the easy-to-update feature of the navigation system.

Liao et al. [20] proposed the integration of the GPS and RFID technologies to develop a system for indoor use in order to address the shortcoming of the GPS system used. This hybrid system receives location data based on GPS and fine tunes the specific location data

with RFID, which was developed to provide walking assistance to users. The study results are expected to facilitate the development of the walking assistance system for the visually impaired individuals and the enhancement of GPS accuracy.

(a) tactile information stand for blind tourists inside a historic building (b) typhlographics in the form of a book showing a tourist facility

Figure 8. Typhlogics in the form of a book that shows tactile information tables and tourist facilities for blind tourists [17].

2.2. Algorithms

2.2.1. YOLO

You Only Look Once v3(YOLOv3), a Darknet-53 network-based object detection algorithm, passes through layers of various sizes and compares them with object characteristics analyzed in the dataset to detect objects [21–23]. This study used YOLOv3 for object detection to identify objects within the line of sight of users. YOLOv3 has undergone several versions of development, making it more accurate than other algorithms [24–27]. In addition, it is fast and specialized for real-time detection as it searches only once, enabling an object detection from images in real time. According to the study of Redmon et al. [23]. YOLOv3 yielded an mAP of 57.9% in a COCO dataset test, demonstrating the high speed and accuracy of the algorithm. Figures 9 and 10 shows the YOLOv3 operating structure and the network structure, respectively. The method detected through the network is shown in Figure 11 and is expressed by Equation (1).

$$\begin{aligned} b_x &= \sigma(t_x) + c_x \\ b_y &= \sigma(t_y) + c_y \\ b_w &= p_w e^{t_w} \\ b_h &= p_h e^{t_h} \end{aligned} \tag{1}$$

2.2.2. Grabcut

The GrabCut algorithm allows more effective object feature classification and ease of use than previous algorithms [28,29], such as Magic Wand, Intelligent Scissors, Bayes Matte, Knockout2, and GraphCut. This algorithm is used to separate the detected objects from the background, exploiting its advantages of high speed and extraction accuracy with only user-specified regions. Through GraphCut-based segmentation, the color values between pixels are calculated. A color model is generated on the basis of the color values of the model, and the foreground and background are separated via segmentation, as shown in Figure 12. After adding a mask to distinguish the foreground and background on the basis of the selection of the user, the separated foreground can be re-extracted, as shown in Figure 13.

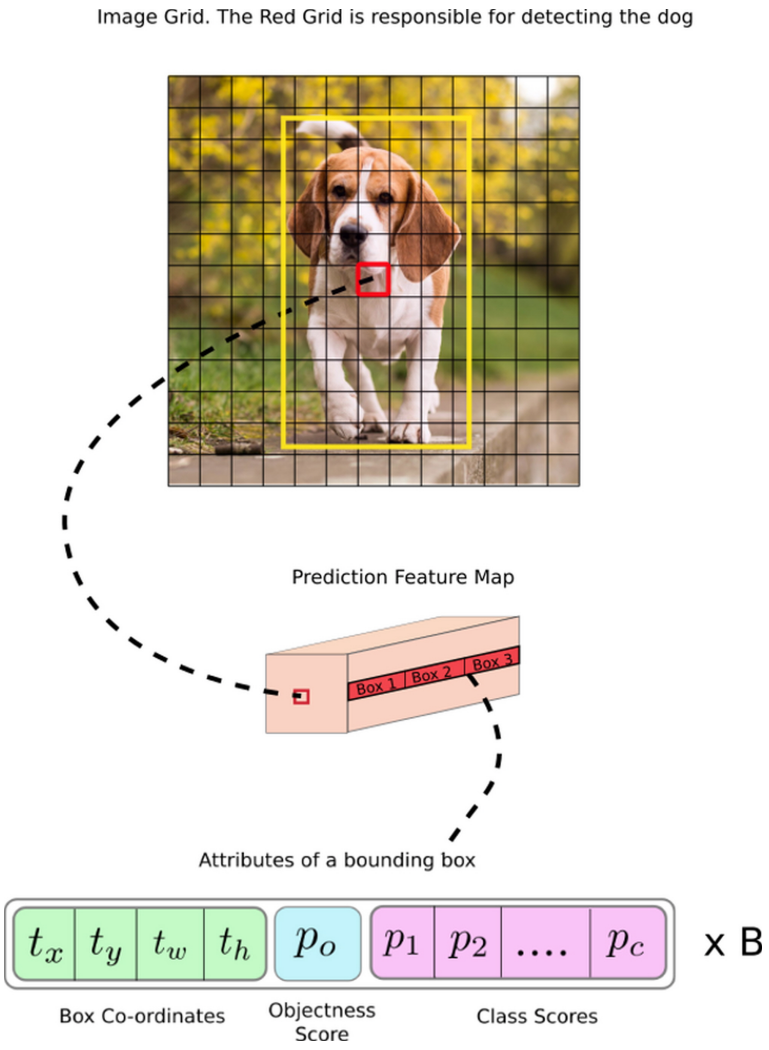


Figure 9. YOLOv3 network detection method [24].

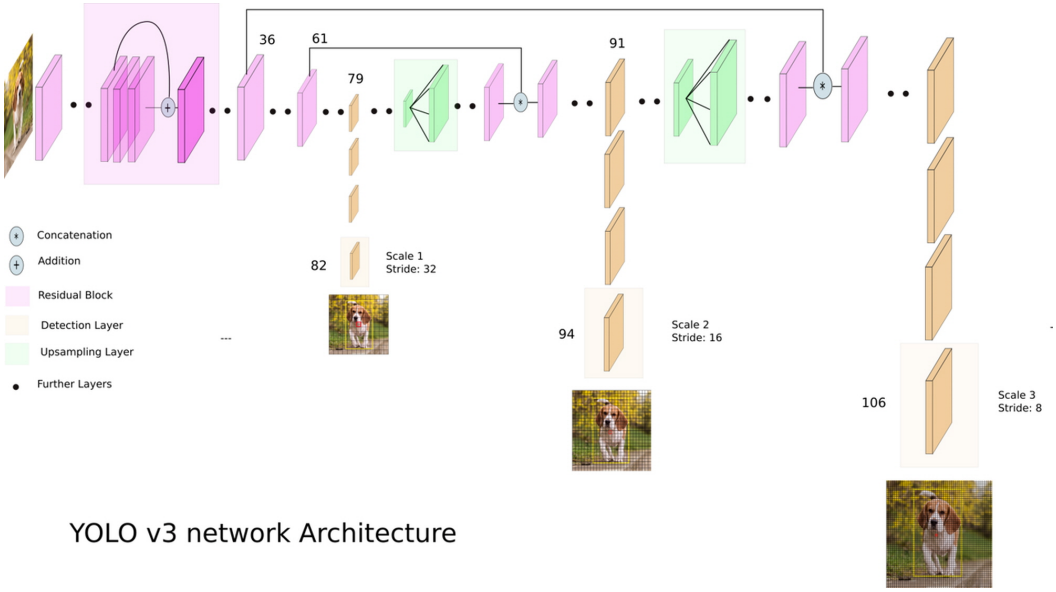


Figure 10. YOLOv3 network architecture [24].

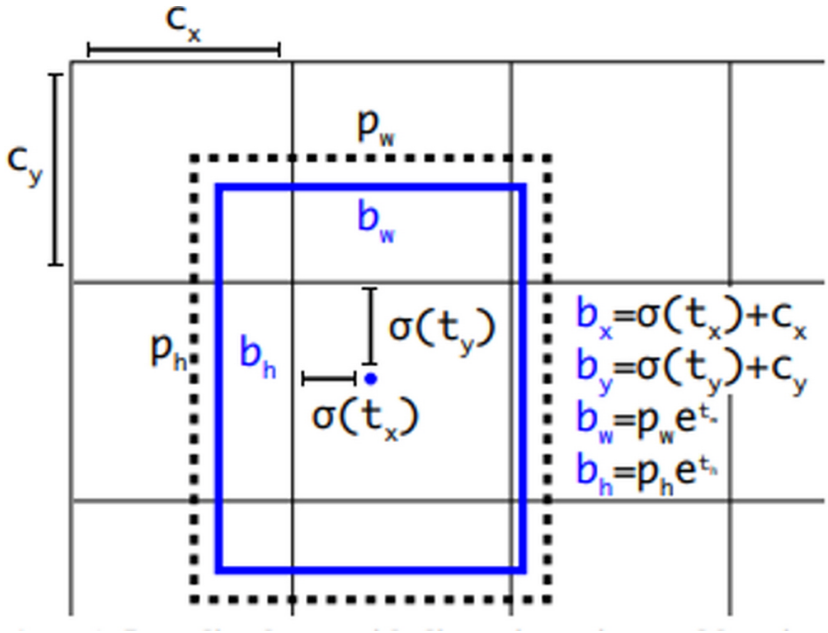


Figure 11. Numerical expression of YOLOv3 object detection [23].

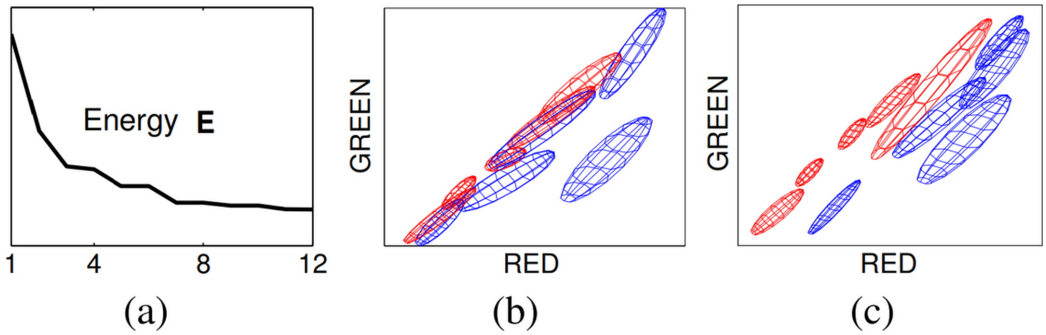


Figure 12. GrabCut principle image-Convergence of iterative minimization. (**a**) The energy E for the llama example converges over 12 iterations. The GMM in RGB colour space (side-view showing R,G) at initialization (**b**) and after convergence (**c**) [28].

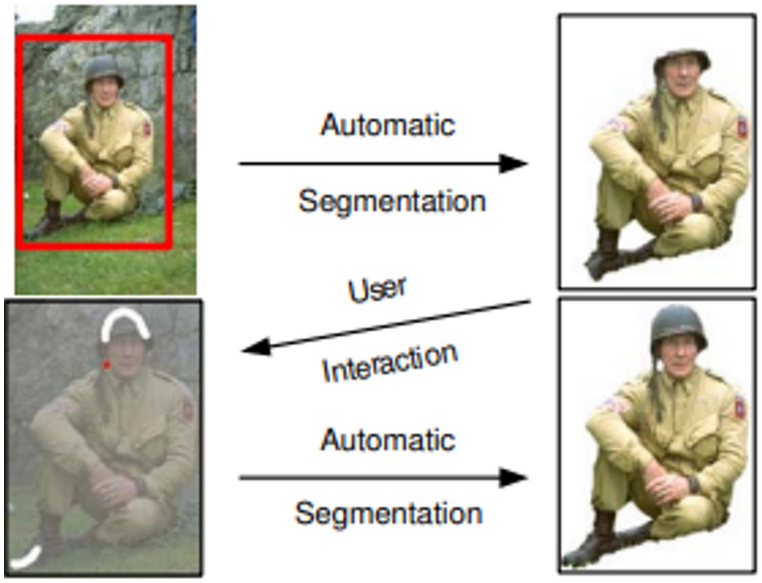


Figure 13. Grabcut example image [28].

2.2.3. Canny

In contrast to Contour, which is a contour line detection algorithm that generates boundary lines based on the height of the boundary detection target [30], Canny identifies the boundary values of the object to generate an outline [31]. In comparison to previous algorithms for generating outlines, Canny is fast and applicable to color images. Therefore, it was used to generate outlines for converting the extracted object to braille. In addition, new criteria were added to prevent it from generating abnormal outlines to achieve a low error rate and stable and improved system performance. Additionally, the criteria of existing algorithms are strengthened, and a parametric closed outline generation technique is provided through numerical optimization. Accordingly, additional criteria were hypothesized, and various equations and operators were used to satisfy the hypotheses. Figure 14 shows the results of this application, indicating its suitability as an outline generation algorithm.

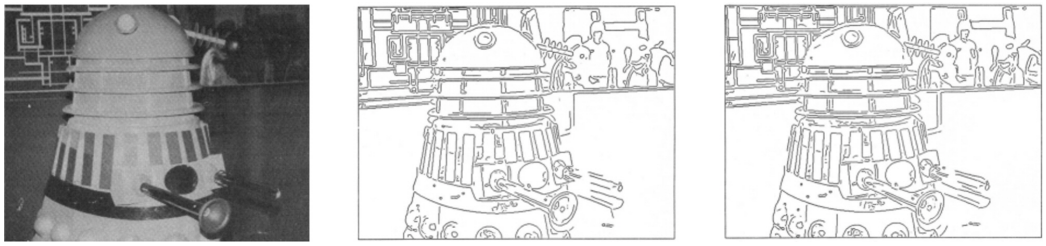


Figure 14. Canny example image [31].

3. System Design and Configuration of Use Environments

The automatic-object-detection-algorithm-based braille conversion system for the living assistance of the visually impaired mainly targets visually impaired people including those with limited sight who typically use braille since the system is fully operated by smartphones. The images of surrounding environment and objects are captured with smart glasses, and the braille images are generated on the braille pads. The relevant objects are captured through smart glasses, and the tactile image is the output on a braille pad. Figure 15 shows the structure of the system, which is operated through a smartphone. To detect objects, it is connected to smart glasses via Bluetooth using the smartphone. The camera screen of the smart glasses and the screen of the desired field of view are confirmed through the smartphone and a shooting request is sent when the smart glasses are connected. When the shooting request reaches the smart glasses, it takes a photo with the built-in camera and sends it to the smartphone. The location and name of the objects in the photo are transmitted through the smart glasses and confirmed via TTS when the system performs object detection at the request of the user. The image is converted to braille, and the braille data are transmitted to the braille pad to allow the user to confirm the shape of the object. Once the transmission is completed, the user can recognize the shape of the object with the tactile image generated through the braille pad.

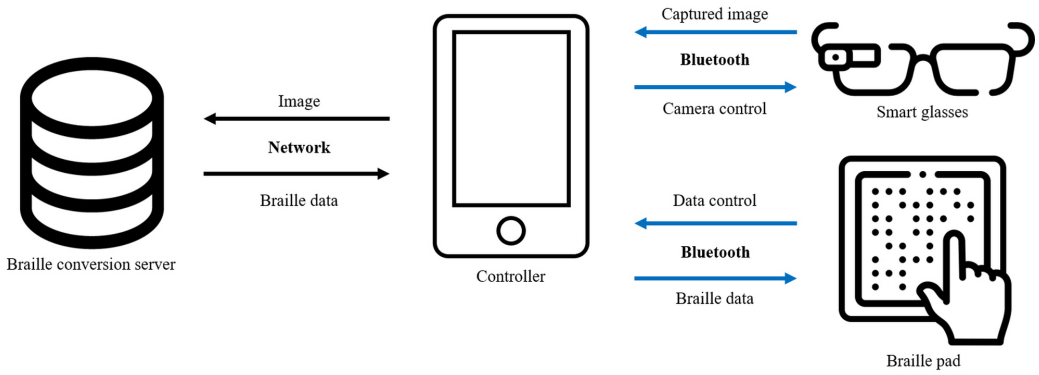


Figure 15. System schematic.

4. System Configuration

Table 1 shows the configuration of the proposed system in five steps: shooting, object detection, object extraction, outline generation, and braille conversion. The algorithms for all steps except shooting are constructed on an integrated server to increase the processing speed and store and use various image data. Each step can be separately executed through a smartphone on the basis of the scope of use and selections of the user. Moreover, only the result data are stored on the smartphone. The data from each step are maintained until the step is executed again. Figure 16 presents the overall process of the system.

Table 1. Requirements of proposed system.

Function	Description
Image shooting	Capture photo of user-specified field of view and generate image
	Transfer to image controller and store
	Receive voice guidance data at user request
Object detection	Learn object images in database defined by system administrator
	Generate object recognition model
	Recognize objects based on image and store result image
	Store analysis result data
	Transmit voice guidance data at user request
Object extraction	Extract objects from image based on data
	Resize and store extracted object images
Outline generation	Preprocess image
	Calculate average color values based on extracted object images
	Generate object outline based on color values
Braille conversion	Analyze generated outline and create braille data
	Analyze resolution of linked braille pad
	Convert data size to braille pad resolution

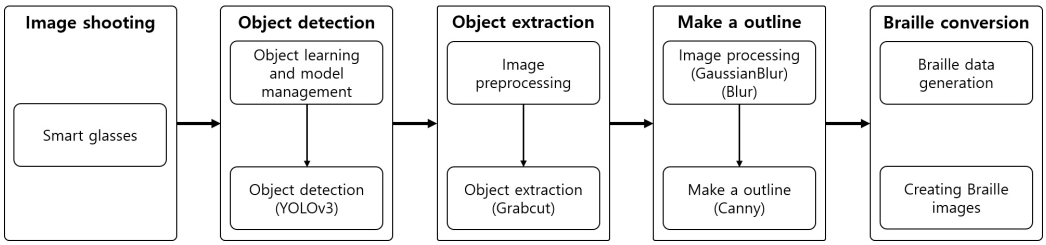


Figure 16. process.

4.1. Object Detection

In the object detection step, the YOLOv3 algorithm was used to detect a variety of objects in real time. Figure 17 shows the results from the application of the system to a real object.

Figure 17. Object detection example [32].

4.2. Object Extraction

The extraction step was configured using Python, and image processing algorithms used were from OpenCV. The objects were extracted using GrabCut after preprocessing the image. Figure 18 shows the structure of the object extraction step.

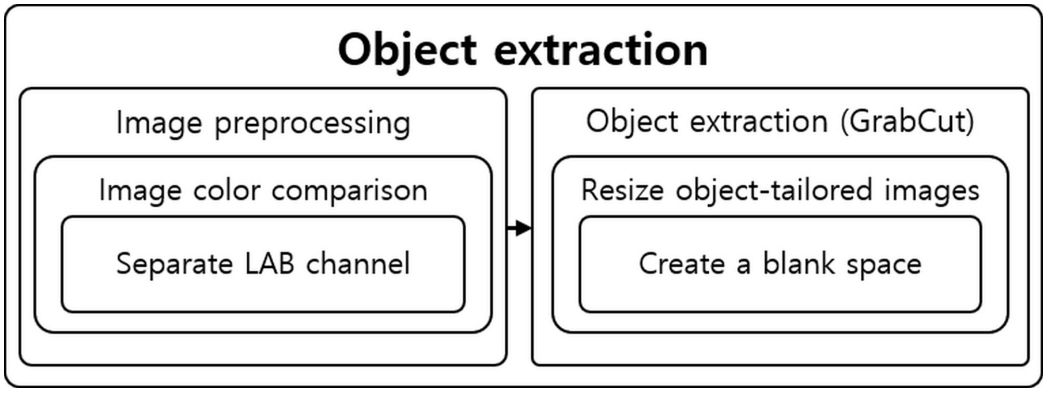


Figure 18. Structural diagram of object extraction steps.

4.2.1. Image Preprocessing

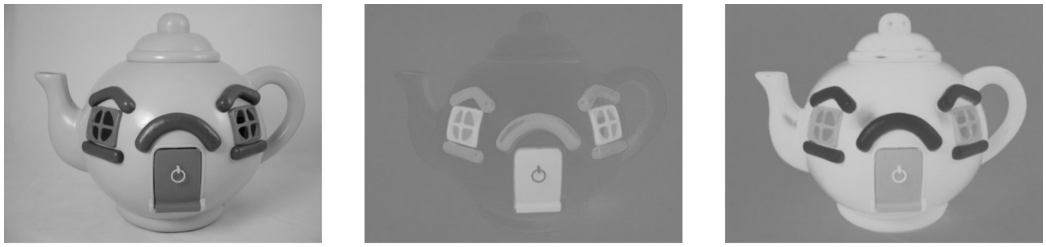
The contrast of the entire image, which refers to the difference in brightness between bright and dark areas in an image, is enhanced to clearly distinguish the colors of the detected image. An image with a small difference in brightness between bright and dark areas has a low contrast value, while an image with a large difference in brightness between bright and dark areas has a high contrast value. The contrast value refers to the contrast ratio. To increase the contrast value, dark areas must be darkened by increasing the color value of the pixels, and bright areas must be brightened by lowering the color values of the pixels.

Although there are various algorithms for increasing contrast value, the most basic technique is to multiply each pixel by a value based on the desired brightness of 1.0 [33,34]. Multiplication techniques are categorized into two methods: multiplying a MAT and using the saturate equation through the clip algorithm. However, they are not suitable for this study because these methods are mainly used on grayscale images to adjust only the brightness values. Instead, we examined algorithms used for colored images. The contrast of colored images is adjusted using a histogram equalization algorithm [35]. In addition, histogram smoothing converts a colored image composed of RGB channels into YCrCb channels and separates them into individual Y, Cr, and Cb channels, respectively, as shown in Figure 19. Y represents the luminance component, while Cr and Cb represent the chrominance components. Histogram equalization is applied to the separated luminance channels to increase the contrast value of the image.

Figure 19. Histogram equalization Example.

Histogram equalization can be applied to an image composed of RGB channels to increase the contrast of the image. It increases the contrast by converting a colored image composed of RGB channels to YCrCb channels and separating them into individual Y, Cr, and Cb channels, respectively. Y represents the luminance component, while Cr and Cb represent the chrominance components. The contrast of color images is increased by applying the histogram equalization in the separated luminance component.

However, histogram equalization adjusts the contrast value of the entire image at once, making the bright areas very bright and dark areas very dark. This results in an unbalanced image overall. The CLAHE algorithm, which separately adjusts the brightness of specific areas in the image, was used to adjust the average brightness while increasing the contrast value [36]. To apply CLAHE, the image is converted to the LAB format and separated into individual channels to separate it into colored and grayscale [37] images. Channel L represents the brightness of the light and is expressed as a black and white image, while channels A and B represent the degree of color. Channel A represents magenta and green, and channel B represents blue and yellow. Moreover, the images are sequentially searched on the basis of the specified grid size, and the contrast value is adjusted to increase the contrast value in channel L and the black and white images. The channels are combined and converted back to the RGB format for other image processing after searching all images and adjusting the contrast value. Using the image from Figure 17, the contrast of an image was increased (Figure 20) through image channel separation, as shown in Figure 21.



(a) Channel L (b) Channel A (c) Channel B

Figure 20. LAB Image by channel.

Figure 21. Contrasted image.

4.2.2. Object Extraction

The stored object location information is imported to extract objects from the image whose contrast was increased in the preprocessing step. Approximately 10 is added to or subtracted from each x and y value in the stored object location information to distinguish the surrounding pixels easily, as shown in Figure 22. GrabCut is used for object extraction. A black background is generated around it when an object is extracted, leaving only the object. In addition, the image size is reduced on the basis of the location information to fit the image size to the object and save it. Figure 23 shows the result of using the GrabCut algorithm.

4.3. Outline Generation

It is hard for users who have difficulty distinguishing objects to recognize objects with large amounts of information at once. Therefore, the tactile image generation was divided into three types depending on the desired type of expression of the user. These three types were "Out," which displays only the outermost part such that the user can recognize the overall shape of the object; "Feature," to ensure that the user can recognize the inner boundaries and form of the object; and "Detail," which displays all information even the text in the object. Figure 24 shows the structure of the outline generation step.

(a) Indicating location information (b) Redefining scope

Figure 22. Object measurement range.

Figure 23. Object extracted image.

4.3.1. Image Processing

In the image processing step for generating the outline, the noise was removed and the colors were averaged. GaussianBlur was performed to remove the noise created by increasing the contrast and other noise [38]. GaussianBlur is used to remove large noise, while averaging [39,40] removes small components and detailed features, such as letters and shapes. Each algorithm was performed with varying degrees of frequency and intensity depending on the outline generation type selected by the user.

In the Out mode, starting from the 7 × 7 kernel and sigma 0, the algorithm was run as it gradually reduces the search size to ensure an iterative and powerful preprocessing and to completely remove noise, features, and information. In the Feature mode, starting from the 5 × 5 kernel and sigma 0, the algorithm was run as it gradually reduces the search size to moderately remove noise and information. On the other hand, in the Detail mode, it

searched with a 3 × 3 kernel and sigma 0 to remove noise while maintaining features and information. Figure 25 shows the image processing results.

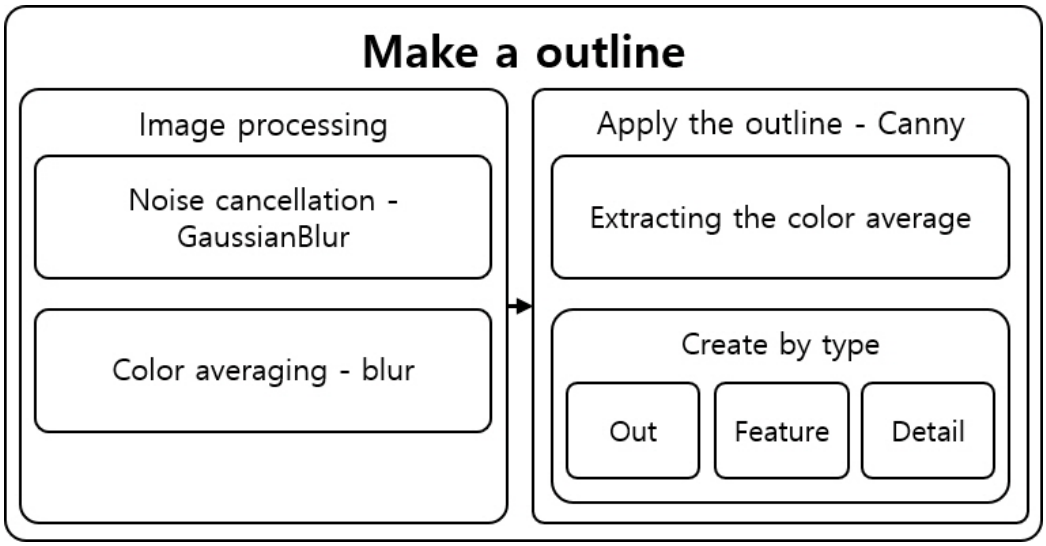


Figure 24. Outline creation step structure.

Figure 25. Image after performing.

4.3.2. Outline Generation

Canny [31] was used because it had a higher speed than Contour [30] although both Contour and Canny yielded similar accuracies for the outline generation algorithm. To generate the outline for each mode, the arrangement average and standard deviation of the images are calculated based on the noise-removed image, and the sum is set to a maximum value, so that the outline generation degree varies depending on the value range and mode. The morphology operations erosion and dilation were used to remove noise and small outlines remaining in the generated outline image. For braille conversion, the thickness was increased three times to confirm the line region, and the generated outlines were stored as individual images according to the mode. The thickness was increased three-fold by repeating the morphology dilation operation [41] three times, and the generated outlines were saved as an individual image on the basis of the mode to clearly define the lines for braille conversion. Figure 26 shows the result of outline generation.

4.4. Braille Conversion

Finally, the braille data were generated in the braille conversion step. For the data size, an image with a horizontal or vertical size of 416 was used as an input in the detection network in YOLOv3. The transformed image was resized on the basis of the detected location information of the object in the object detection process. Moreover, the data size was converted through braille data resizing on the basis of the received braille pad resolution when the braille pad was connected, ensuring that the output braille fitted the braille pad.

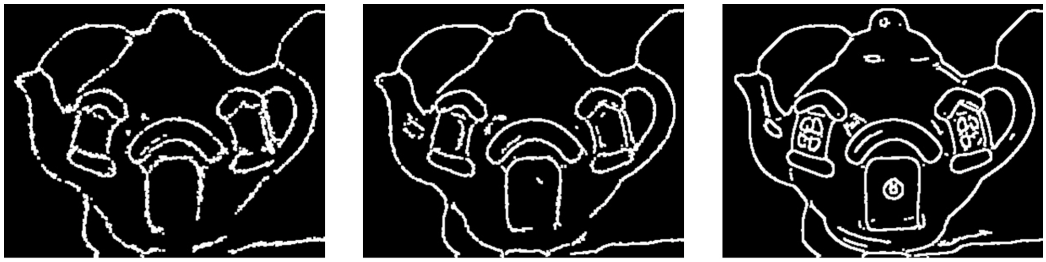


Figure 26. Outline image—Step-by-step image completed up to thickness increase.

n × n Comparison Conversion

At this stage, the outline images generated through Python were imported and converted to braille data to create braille images. Colored values in which the color and brightness can be identified were searched via array comparison because images in Python are expressed as an array. A two-dimensional array of the same size as the image was generated to perform the search. The image was searched in a 5 × 5 pixel neighborhood, and it was checked whether there are color data in the center pixel (>0), as shown in Figure 27. A value of 1 was stored in the same location as the generated two-dimensional array if there are color data. An array was finally generated by searching the entire image, which is stored for transmission to the braille pad. The tactile image was generated by the same technique; the image was searched, and a circle was created in areas with a value. Figure 28 shows the results of braille transformation through comparative transformation.

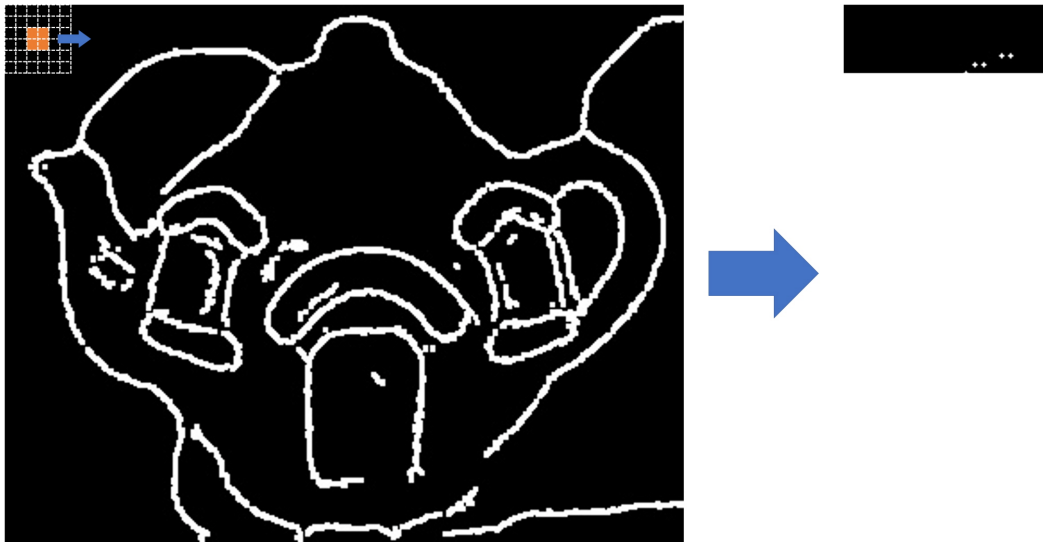


Figure 27. Example of braille conversion process.

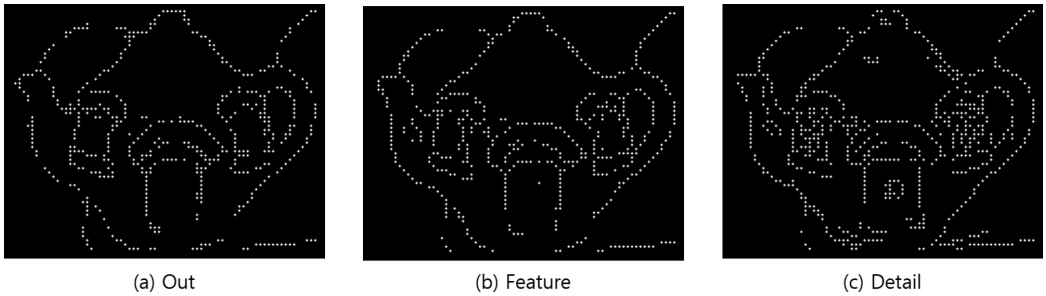


Figure 28. Generated braille image.

5. Experiment and Evaluation

We evaluated the accuracy and usefulness of the tactile image generated by the proposed system. To evaluate the accuracy, the expected result images and the system result images for a variety of objects were compared. On the other hand, to evaluate the usefulness, the execution time of the system was calculated using photos in diverse situations that can be confirmed in real life, which verified the applicability of the system in real life.

5.1. Experiment

5.1.1. Object Data Generation

The highly well-known and stable Microsoft COCO dataset [42] was used as the basic dataset because the dataset was required for object detection through YOLOv3. Table 2 lists the selected objects. Additionally, based on the COCO dataset object list, objects that give visually impaired people discomfort were added according to survey results, thus forming a dataset with 100 types of objects. The survey was conducted among visually impaired people in Korea at a welfare center. Table 3 summarizes the results.

5.1.2. Accuracy Evaluation

To select the objects for the evaluation criteria, the objects that the visually impaired frequently use or encounter in real life were categorized into the following: (1) "indoors" and "outdoors" and (2) based on their sizes (i.e., large, medium, and small), resulting in a total of six objects. The following size criteria were applied: objects difficult to hold in the hands were classified as large, objects that can be held with two hands as medium, and objects that can be held with one hand as small. For the objects that are most frequently encountered outdoors, "car" was selected for large, "fire hydrant" for medium, and "traffic cone" for small. On the other hand, "closet" was selected for large, "chair" for medium, and "comb" for small for the objects that are most frequently used indoors. Figures 29 and 30 show the comparison between the expected data and actual object results. The actual results were compared with [43,44] the braille for the "Detail" mode to verify the expression of details in the images.

5.1.3. Usefulness Evaluation

To evaluate the usefulness, based on the three photos with the themes of "walking," "eating," and "washing face," the conversion time in each step was measured and averaged, and the identified objects were compared with [43,44] the detected object list. Only the name and location value of the object closest to the user were used when there were duplicate objects in the braille conversion step, thus performing braille conversion without any duplicate objects. Figure 31 shows the photos used for the evaluation, converted photos, and detected objects list, with conversion times of 5.8, 4.5, and 7.4 s, respectively.

5.2. Overall Evaluation

The main object was compared with the expected generated data to evaluate the accuracy of the tactile image. We verified the amount of time needed for conversion to evaluate the usefulness of the system.

In the accuracy evaluation, the expected result image was visually compared with the resulting image of the system, and the accuracy of the generated tactile image was measured. The results showed that the final image has an average accuracy of 85% which is similar to that of the expected image.

In the usefulness evaluation, the list of detected objects was compared and the conversion time was measured on the basis of the photos of three situations that users can encounter in real life. For the objects detected in photos of real-life situations, the results indicated an accuracy of approximately >90%. By excluding duplicate objects, the average time needed to convert the objects was less than 6.6 s, exhibiting that it can be quickly used in real life.

Table 2. COCO dataset object list [42].

Person	Backpack	Umbrella	Handbag	Tie	Suitcase	Bicycle	Car	Motorcycle	Airplane
Bus	Train	Truck	Boat	Traffic light	Fire hydrant	Stop sign	Parking meter	Bench	Bird
Cat	Dog	Goose	Sheep	Cow	Elephant	Bear	Zebra	Giraffe	Frisbee
Skis	Snowboard	Sports ball	Kite	Baseball bat	Baseball glove	Skateboard	Surfboard	tennis racket	Bottle
Wine glass	Cup	Fork	Knife	Spoon	Bowl	Banana	Apple	Sandwich	Orange
Broccoli	Carrot	Hot dog	Pizza	Donut	Cake	Chair	Couch	Potted plant	Bed
Dining table	Toilet	TV	Laptop	Mouse	Remote	Keyboard	Cell phone	Microwave	Oven
Toaster	Sink	Refrigerator	Book	Clock	Vase	Scissors	Teddy bear	Hair drier	Toothbrush

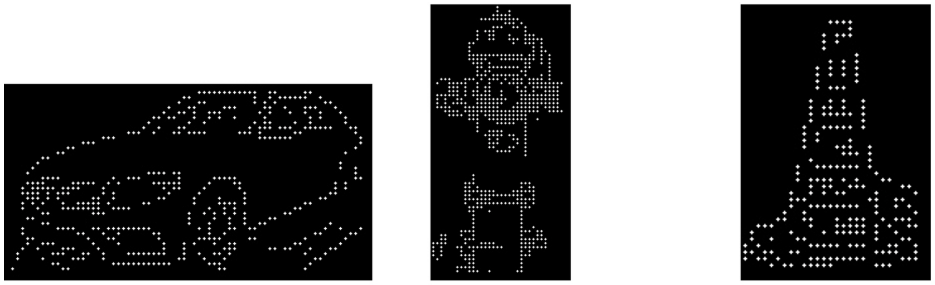
Table 3. List of selected objects.

Person	Backpack	Umbrella	Handbag	Tie	Suitcase	Bicycle	Car	Motorcycle	Airplane
Bus	Train	Truck	Traffic light	Fire hydrant	Subway	Bench	Bird	Cat	Dog
Sports ball	Skateboard	Bottle	Wind glass	Cup	Fork	Knife	Spoon	Bowl	Chair
Tissu	Potted plant	Bed	Dining table	Toilet	TV	Laptop	Mouse	Remote	Keyboard
Cell phone	Microwave	Sink	Refrigerator	Book	Clock	Pillow	Scissors	Toothbrush	Toothpaste
Hair drier	Braille pad	Tree	Street lamp	Utility pole	Manhole	Vending machine	Elevator	Standing board	Escalator
Shampoo	Conditioner	Lottion	Stair	Traffic cone	Bollard	Radio	Desk	Whellchair	Eletric rice cooker
Gas cooker	Closet	Washing machine	Teapot	Electric fan	Comb	Bookmark	Soap	Glasses	Key
Shoes	Shower	Tumbler	Walking stick	Plate	Pencil	Electric kettle	Pen	Eraser	Earphones
Towel	Chopsticks	Meat	Fish	Hat	Rice	Kimchi	Bread	Cushin	Mattress

This system can output tactile images generated on the basis of braille data of objects with a shape similar to those of real-life objects, yielding significant results.

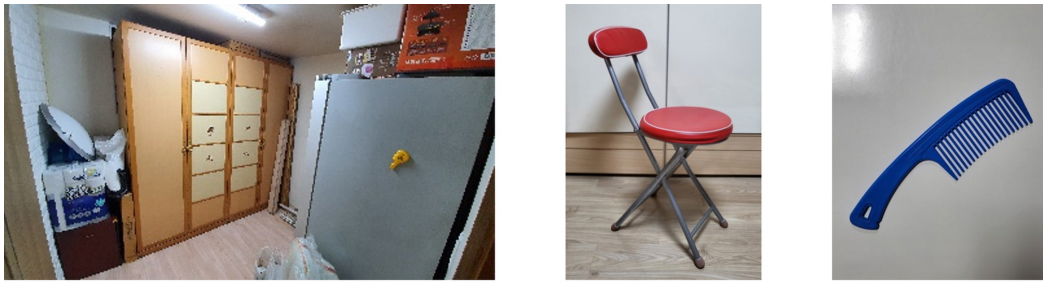
Ten visually impaired individuals were satisfied with the performance of the assistance system. Moreover, they preferred the Out type, which simplifies the tactile information in a straightforward manner, over the Detail type, which converts the real objects of complex composition.

(a) Actual image

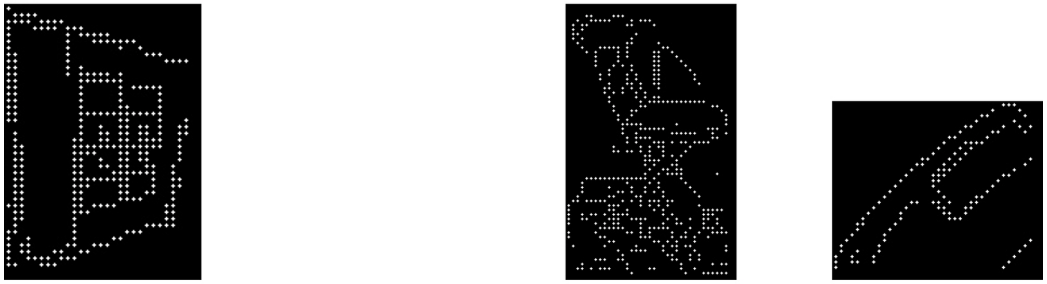


(b) Expected result

Figure 29. Comparison of expected and actual data(Outdoors).



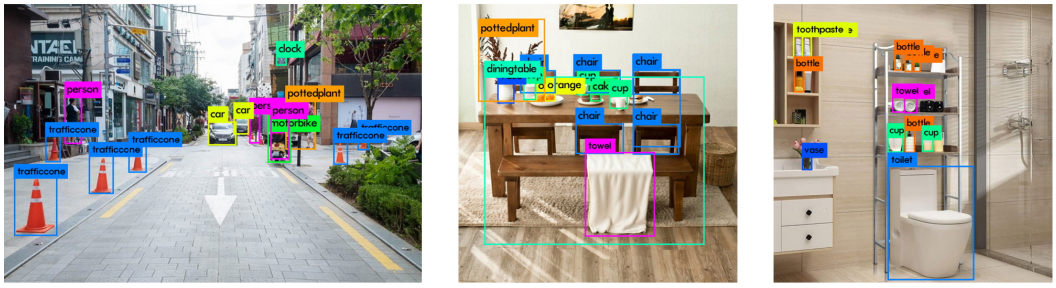
(a) Actual image



(b) Expected result

Figure 30. Comparison of expected and actual data(Indoors).

(a) Panoramic photos in action



traffic cone, car, motorcycle, plant, person

towel, cup, plate, flowerpot, chair, dining table, orange, cake

toilet bowl, towel, potted plant, cup, bottle, toothpaste

(b) Detection result

Figure 31. As a result of applying it to real life photos.

6. Conclusions and Discussion

6.1. Conclusions

The proposed system was designed to inform visually impaired people about the types of obstacles in their field of view and to help them recognize their shapes. The system used an AI algorithm with high processing speed to quickly guide the user and integrated a simple image processing algorithm to provide tactile images in a short time. This study proposes a new and simple type of assistive device for visually impaired people who usually use braille, including people with limited sight. However, new algorithms or the latest technologies were not applied in the proposed system. The proposed braille conversion algorithm yielded an accuracy of 85% in relation to the expected result, demonstrating its usefulness. By excluding duplicate objects, approximately 12 out of 13 objects that can be confirmed in real life were detected on average. In addition, the conversion took an average of 6.6 s, indicating that the system is sufficient for use in real life.

6.2. Discussion

This study proposes a living assistance system that is applicable both indoors and outdoors by integrating object recognition, object extraction, outline generation, and braille conversion algorithms. According to the experiments and evaluations, we found that the system developed on the basis of the database tailor-made to the needs of visually impaired people (includes people with limited sight), who usually use braille, was useful.

However, some limitations of this study include the object extraction results obtained through GrabCut using the coordinates of the detected objects with YOLOv3 did not match with the real object. Moreover, some images other than the object image are left, indicating an inaccuracy in the braille conversion.

Therefore, we plan to perform primary development research to further improve the accuracy of the system and to generate and apply YOLOv3-based object masks although a more advanced system may require additional conversion time. Furthermore, we plan to conduct secondary development research to convert detected objects to icons and reflect the areas of improvement found from tests.

Author Contributions: D.L. designed, contributed to system model and implemented testbed. J.C. contributed to paper review and formatting. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2019R1F1A105775713) and This work was supported by the Gachon University research fund of 2020 (GCU-202008460007).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Publicly available datasets were analyzed in this study. COCO dataset can be found here: https://cocodataset.org/ (accessed on 30 September 2021).

Acknowledgments: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2019R1F1A105775713) and This work was supported by the Gachon University research fund of 2020 (GCU-202008460007).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Korea Institute for Health and Social Affairs. Cause and Timing of Disability. 2019. Available online: https://kosis.kr/statHtml/statHtml.do?orgId=331&tblId=DT_33109_F37&conn_path=I2 (accessed on 30 September 2021).
2. The Online Database of Health Reporting (GBE). The Information System of the Federal Health Report and Statement of the Federal Republic of Germany. 2015. Available online: https://www.gbe-bund.de/gbe/ (accessed on 1 October 2021).
3. World Health Organization (WHO); World Bank. *World Report on Disability 2011*. 2011. Available online: https://www.who.int/disabilities/world_report/2011/report.pdf (accessed on 2 October 2021).
4. World Health Organization (WHO). Blindness and Vision Impairment. 2021. Available online: https://www.who.int/en/news-room/fact-sheets/detail/blindness-and-visual-impairment (accessed on 14 October 2021).
5. Ministry of Health and Welfare. The Number of Registered Disabled Persons by Type of Disability and Gender Nationwide. 2020. Available online: https://kosis.kr/statHtml/statHtml.do?orgId=117&tblId=DT_11761_N001&conn_path=I2 (accessed on 14 October 2021).
6. German Federal Statistical Office. People with Severe Disabilities with ID (Absolute and 100 per Person). (Population of 1000). Features: Years, Region, Type of Disability, Degree of Disability. 2019. Available online: https://www.gbe-bund.de/gbe/pkg_isgbe5.prc_menu_olap?p_uid=gast&p_aid=21134557&p_sprache=D&p_help=0&p_indnr=218&p_indsp=&p_ityp=H&p_fid= (accessed on 15 October 2021).
7. Kostopoulos, K.; Moustakas, K.; Tzovaras, D.; Nikolakis, G. Haptic Access to Conventional 2D Maps for the Visually Impaired. In Proceedings of the 2007 3DTV Conference, Kos, Greece, 7–9 May 2007; pp. 1–4. [CrossRef]
8. Zeng, L.; Miao, M.; Weber, G. Interactive audio-haptic map explorer on a tactile display. *Interact. Comput.* **2015**, *27*, 413–429. [CrossRef]
9. Krufka, S.E.; Barner, K.E.; Aysal, T.C. Visual to tactile conversion of vector graphics. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2007**, *15*, 310–321. [CrossRef] [PubMed]
10. Krufka, S.E.; Barner, K.E. A user study on tactile graphic generation methods. *Behav. Inf. Technol.* **2006**, *25*, 297–311. [CrossRef]
11. Kim, H.J.; Kim, Y.C.; Park, C.J.; Oh, S.J.; Lee, B.J. Auto Braille Translator using Matlab. *J. Korea Inst. Electron. Commun. Sci.* **2017**, *12*, 691–700.
12. Jiwan Lee, J.A.; Lee, K.Y. Development of a raspberry Pi-based banknote recognition system for the visually impaired. *J. Soc.-Bus. Stud.* **2018**, *23*, 21–31. [CrossRef]
13. Hahn, M.E.; Mueller, C.M.; Gorlewicz, J.L. The Comprehension of STEM Graphics via a Multisensory Tablet Electronic Device by Students with Visual Impairments. *J. Vis. Impair. Blind.* **2019**, *113*, 404–418. [CrossRef]
14. Kim, S.; Park, E.S.; Ryu, E.S. Multimedia vision for the visually impaired through 2d multiarray braille display. *Appl. Sci.* **2019**, *9*, 878. [CrossRef]
15. Kim, S.; Yeongil Ryu, J.C.; Ryu, E.S. Towards Tangible Vision for the Visually Impaired through 2D Multiarray Braille Display. *Sensors* **2019**, *19*, 5319. [CrossRef]
16. Prescher, D.; Bornschein, J.; Köhlmann, W.; Weber, G. Touching graphical applications: Bimanual tactile interaction on the HyperBraille pin-matrix display. *Univers. Access Inf. Soc.* **2018**, *17*, 391–409. [CrossRef]
17. Kłopotowska, A.; Magdziak, M. Tactile Architectural Drawings—Practical Application and Potential of Architectural Typhlographics. *Sustainability* **2021**, *13*, 6216. [CrossRef]
18. H.Morad, A. GPS Talking For Blind People. *J. Emerg. Technol. Web Intell.* **2010**, *2*, 239–243. [CrossRef]
19. Fernandes, H.; Filipe, V.; Costa, P.; Barroso, J. Location based Services for the Blind Supported by RFID Technology. *Procedia Comput. Sci.* **2014**, *27*, 2–8. [CrossRef]

20. Liao, C.; Choe, P.; Wu, T.; Tong, Y.; Dai, C.; Liu, Y. RFID-Based Road Guiding Cane System for the Visually Impaired. In *Cross-Cultural Design. Methods, Practice, and Case Studies*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 86–93.
21. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
22. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
23. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
24. Kathuria, A. What Is New in YOLO v3? 2018. Available online: https://towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b (accessed on 6 September 2021).
25. Lee, S.; Lee, G.; Ko, J.; Lee, S.; Yoo, W. Recent Trends of Object and Scene Recognition Technologies for Mobile/Embedded Devices. *Electron. Telecommun. Trends* **2019**, *34*, 133–144.
26. Poudel, S.; Kim, Y.J.; Vo, D.M.; Lee, S.W. Colorectal disease classification using efficiently scaled dilation in convolutional neural network. *IEEE Access* **2020**, *8*, 99227–99238. [CrossRef]
27. Siddiqui, Z.A.; Park, U.; Lee, S.W.; Jung, N.J.; Choi, M.; Lim, C.; Seo, J.H. Robust powerline equipment inspection system based on a convolutional neural network. *Sensors* **2018**, *18*, 3837. [CrossRef]
28. Rother, C.; Kolmogorov, V.; Blake, A. "GrabCut" interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph. TOG* **2004**, *23*, 309–314. [CrossRef]
29. OpenCV. Interactive Foreground Extraction Using GrabCut Algorithm. Available online: https://docs.opencv.org/master/d8/d83/tutorial_py_grabcut.html (accessed on 7 September 2021).
30. Ghuneim, A.G. Contour Tracing. 2000. Available online: http://www.imageprocessingplace.com/downloads_V3/root_downloads/tutorials/contour_tracing_Abeer_George_Ghuneim/author.html (accessed on 7 September 2021).
31. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *PAMI-8*, 679–698. [CrossRef]
32. Berry, S. Big Yellow Teapot. Available online: https://www.flickr.com/photos/unloveable/2388661262 (accessed on 7 September 2021).
33. Gonzalez, R.C.; Woods, R.E. *Digital Image Processing*; Pearson/Prentice Hall: Hoboken, NJ, USA, 2008.
34. Laughlin, S. A Simple Coding Procedure Enhances a Neuron's Information Capacity. *Z. Für Naturforschung C* **1981**, *36*, 910–912. [CrossRef]
35. Hummel, R.A. Image Enhancement by Histogram transformation. *Comput. Graph. Image Process.* **1975**, *6*, 184–195. [CrossRef]
36. Pizer, S.M.; Amburn, E.P.; Austin, J.D.; Cromartie, R.; Geselowitz, A.; Greer, T.; ter Haar Romeny, B.; Zimmerman, J.B.; Zuiderveld, K. Adaptive histogram equalization and its variations. *Comput. Vis. Graph. Image Process.* **1987**, *39*, 355–368. [CrossRef]
37. International Commission on Illumination. *Colorimetry*; CIE Technical Report; Commission Internationale de l'Eclairage: Vienna, Austria, 2004.
38. Haddad, R.; Akansu, A. A class of fast Gaussian binomial filters for speech and image processing. *IEEE Trans. Signal Process.* **1991**, *39*, 723–727. [CrossRef]
39. Kalman, R.E. A New Approach to Linear Filtering and Prediction Problems. *Trans. ASME–J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
40. Gonzalez, R.; Wintz, P. *Digital Image Processing*, 2nd ed.; Addison-Wesley: Boston, MA, USA, 1987.
41. OpenCV. Eroding and Dilating. Available online: https://docs.opencv.org/4.x/db/df6/tutorial_erosion_dilatation.html (accessed on 8 September 2021).
42. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. *arXiv* **2014**, arXiv:1405.0312.
43. Park, J.H.; Whangbo, T.K.; Kim, K.J. A novel image identifier generation method using luminance and location. *Wirel. Pers. Commun.* **2017**, *94*, 99–115. [CrossRef]
44. Wang, H.; Li, Z.; Li, Y.; Gupta, B.; Choi, C. Visual saliency guided complex image retrieval. *Pattern Recognit. Lett.* **2020**, *130*, 64–72. [CrossRef]

Smart Materials Detection Using Computer Vision

Conference Paper · September 2019

CITATIONS
5

READS
1,078

2 authors:

Mustafa Bilgin
University of Duisburg-Essen
34 PUBLICATIONS **42** CITATIONS
SEE PROFILE

Johannes Backhaus
Bergische Universität Wuppertal
25 PUBLICATIONS **47** CITATIONS
SEE PROFILE

Smart Materials Detection Using Computer Vision

Mustafa Bilgin, Johannes Backhaus

University of Wuppertal
Rainer-Gruenter-Straße 21
42119 Wuppertal, Germany

E-mail: bilgin@uni-wuppertal.de; jbackhaus@uni-wuppertal.de

Abstract
Smart materials detect autonomously changings of environmental conditions such as temperature (thermochromism), water / humidity (hydrochromism), UV/Vis light (photochromism), mechanical stress (piezochromism), acids (pH-chromic) and more. The irreversible colour changing of smart materials can be used to "translate" the particular colour information into the degree of contamination, such as a printable sensor. Integrated into a smart code, besides static data (2D) also dynamic sensor data (3D) can be stored. The aim of this study is the computer vision based image detection and interpretation of a smart material surface. The particular colour information will be transformed into the degree of contamination. The limits of feasibility will be examined on a laboratory scale and comparison with reference colour values.

Keywords: smart materials, neural networks, computer vision, smart code, piezoelectric inkjet

1. Introduction and Background

Smart Materials can show a hydrochromic (water/humidity), photochromic (UV/Vis light), thermochromic (temperature), piezochromic (pressure) and other behaviours. They gradually can shift between two states by changing their colour. Photochromic compounds show a light detecting behaviour. If the surface of a photochromic compound is exposed due to UV-light, it changes its colour gradually from one colour state to another colour state. The pH-chromic compound is used to determine the acid or base capacity of a substance. This e.g. can be used to monitor the freshness of foodstuffs. These materials are autonomous and independent of electrical circuits and electrical sources. A camera system can optically read the individual smart materials surfaces. Thus, by the recorded colour values can be concluded with a change in colour in terms of the degree of contamination. This colour changing behaviour can be used to develop printable sensors (Bilgin & Backhaus, 2017, a). That way, it is possible to store dynamic information (Smart Materials) about possible deviations and static consumer or product information inside a smart code. By means of computer vision, it is possible to automate tasks by processing, analysing, interpreting, and manipulating digital image information. This paper shows functions for recording a smart code and its subareas.

For evaluating a smart code, the following steps are necessary: Most of the code readers operate with a binarisation initiated by a prior grayscale transformation. This reduces the data volume with regard to short processing time. - It is also helpful to crop the image to a smaller size without the data getting lost - However, a grayscale transformation is not adequate for coloured sensor data as meaningful data can be lost. In this way, image files created in RGB can be processed in different colour conversions such as L*a*b*, which can be used as a reference colour system. It is important to use these colour systems to evaluate or calibrate the colours. Being JPG files, camera images often do not contain the device's profile (ICC), but either one of the standard profiles sRGB or Adobe RGB. These determinants must be controlled. The L*a*b* colour space contains all colours independent of any device. It therefore allows lossless conversion of colour information from one colour system to another, from one device type to another. This provides a subsequent interpretation of the colour information in statements about the degree of contamination. However, in order to correct the image information, the use of morphological operators can be useful. A further aspect is erosion, which causes false classifications to disappear due to erosion. These can be connected image areas. On the other hand, dilatation can be used to compress or strengthen partial areas in a pixel group, to fill holes, to close cracks. The combination of smart

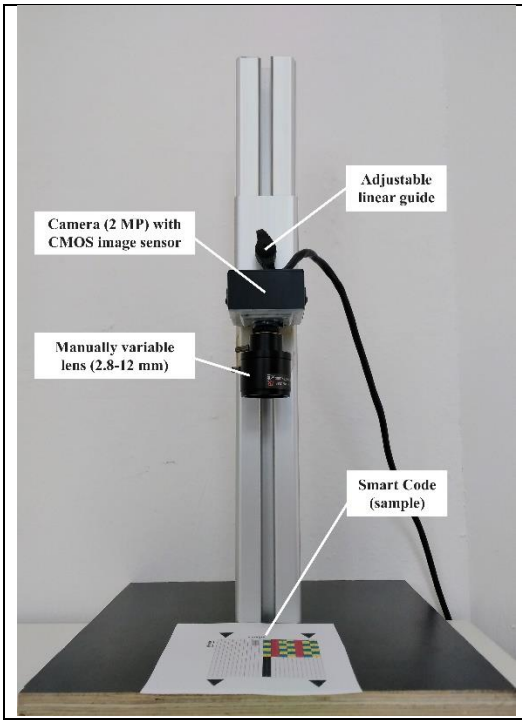
materials, smart devices and smart codes will allow the construction of an Internet of Things (IoT) in which all components communicate in an autonomous network to perform predetermined tasks, provide consumer information, collect feedback on their products and provide information on critical environmental impacts throughout the transport process (Ashton, 2009).

Objective of code development:
The Smart Code in this work is to be divided into three main areas: The static area for text content, the finder pattern area for identifying and aligning the code and the dynamic area for smart materials (printable sensors). The special feature of the static area is that the Braille font is to be used. With the help of the smartphone camera, visually impaired people or the unimpaired can read out the content of the code or touch it manually. In addition, the dynamic range in the smart code can be optically read out and results on the respective state of the smart materials can be output e.g. auditory.

2. Materials and Methods

2.1 Instruments

The smart code samples (containing smart materials) were printed with a piezoelectric inkjet printer (Epson WorkForce WF-3620). Technical parameters: Print Head: PrecisionCore; Thin Film Piezo element: 1/1000mm; Droplet Size: 2.8 pl (range of 1.5 – 32.5 picoliters); Nozzle Configuration: 800 Nozzles Black (K), 256 Nozzles per Colour (CMY); Printing Resolution: 4,800 x 2,400 DPI.



Experimental assembly
Technical parameters: The smart codes were captured with a 2.07 MP camera (ELP-USBFHD01M-BFV-D) with CMOS image sensor and a manually variable 2.8-12 mm lens - the focus of the camera is manually adjustable. The camera has an image resolution of 1080P (1920 × 1080 pixels).
The experimental setup is based on an adjustable linear guide. Thus, the camera can be moved vertically. In this publication, however, the lighting conditions are uncontrolled. However, the construction can be placed in a light chamber in order to realize the evaluation under a reference daylight Illuminant (e.g. D50 standard light). The recording of the samples is done in real time and under real conditions.

For reference purposes, samples were measured with a spectral densitometer (TECHKON SpectroDens) to analyse their characteristic RGB and CIEL*a*b values and to compare them with values detected and analysed by the computer vision. Technical parameters: polarising filter: off; type of light: D50, 2° standard observer; diameter of measuring orifice: 3 mm. This paper uses the OpenCV (Open Source Computer Vision Library), an open source computer vision and machine-learning library, for image processing. The algorithms can be used to recognize faces, identify objects, classify human actions in videos, track camera movements, track-moving objects, recognize scenes or, as in this case, detect smart codes and evaluate partial areas. The programs for analysing the smart codes were programmed in Python. For this purpose, other libraries like NumPy were used. NumPy extends Python with functions for scientific computing and numerical computation (OpenCV, 2019).

2.2 Standardization

All experiments were carried out under controlled laboratory conditions - reproducibility was ensured by an air conditioning system and deviations were recorded in protocols. Temperature: 20 °C (+/- 1 ° C); relative humidity: 55% (+/- 1%) were controlled.

2.3 Materials

All materials, which were used for this research, are listed in Table 1. To ensure the reproducibility, all experiments are based on standardised substrates Inapa tecno. The smart material used for the experiments is photosensitive Prussian blue, described in (Bilgin & Backhaus, 2018).

Substrates for printing	Inapa tecno, oxygen pure high-white recycled paper, Format: 210 x 297 mm (A4), Grammage: 80 g/m²
Dye	Photosensitive Prussian blue, CAS Number: 14038-43-8, Chemical formula: $C_{18}Fe_7N_{18}$, Molar mass: 859.24 g·mol^{-1}
Water-based base ink	E24, Octopus Fluids GmbH & Co KG, Colour: colourless, pH: 7,86, Conductivity: (mS/cm): <5, Viscosity (mPa·s): 3,00

Table 1 *Materials*

2.4 Test Chart

The initial intention was to modify a standard QR code and design dynamic areas (for smart materials) within the code. However, this idea was dropped because the QR codes are designed only for static data. Dynamic (i.e. information changing) areas avoid decoding the QR code. Consequently, a simple smart code containing static and dynamic areas was designed and prototyped in laboratory scale (Fig.1) The code is divided into three sections. The first area (red framed) shows three finder patterns (squares) in three corners. They are required to identify the smart code and determine the inner coordinate system of the code. The second area (grey hatched) is the static data area in which 54 characters of text information can be placed. The third area (yellow framed) is the dynamic area that consists of several cyan, magenta and yellow caches which here serve as placeholders for three different smart materials.

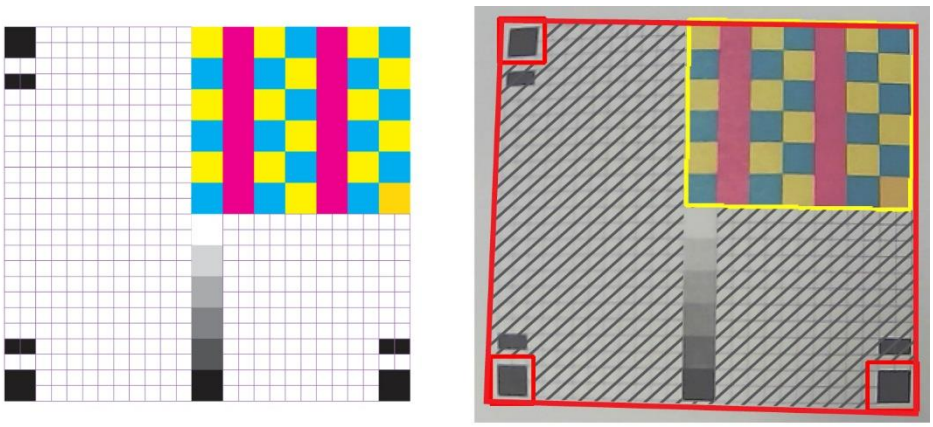


Fig. 1 The test chart shows a smart code prototype.

3. Results and Discussion

3.1 Static Data: Encoding and Decoding

There are many methods to encode information. Therefore, there are many coding schemes (Morse characters, nautical signal flags, semaphore, binary, numbers to letters conversion and more). However, this paper focuses on Braille (Fig.2). It usually consists of six embossed dots, but it was later adapted to computer language and consists of 8-point characters (International Standard ISO/IEC 10646 - Unicode) used by visually impaired people. This scheme allows 256 different characters to be displayed. A special requirement in the coding of the static content is that visually impaired people can read the content

manually by touching it. However, the static data should also be readable and converted into an alphanumeric text (Latin alphabet) for persons who do not master the Braille writing process. - This can be done with the smartphone camera. The dynamic part of the code is also machine-readable and is evaluated by computer vision.

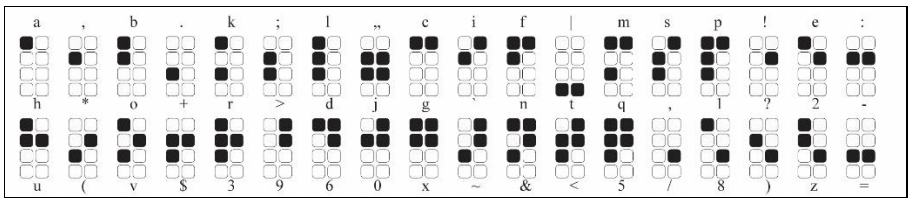


Fig. 2 Unicode 8-dot Braille characters

A usual scanner recognizes an image pixel by pixel. Here, a function is used that helps to identify points, whose dimensions, distances and respective colour information are required. The Hough Transformation (Hough and Paul, 1962) is a method for detecting straight lines, circles (such as the Braille points) or any other parameterizable geometric shapes in a binary black-and-white image. In the following example (Fig. 3) the Hough Circle method was applied. The result for the recognition of a Braille line and the application of the Hugh method must be prepared: Initially, an 8-bit image is converted to grayscalein order to reduce the entire image information to essential information. For the recognition of Braille points we refer to the OpenCV function cv2.HoughCircles in corresponds to Yuen et al. (1990).

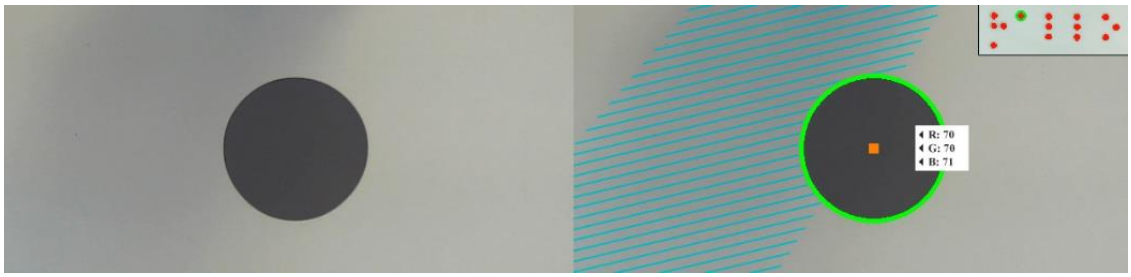


Fig. 3 Circle Hough Transform

The static area of the code is very important for a calibration of the camera in order to avoid external optical influences when analysing the colour of the sensitive dynamic code. By detecting the black points from the Braille scheme and the colour information of the white background, it is possible to use these information for a calibration of the camera. Each point of the Braille scheme and its adjacent background can be used to calibrate both the exposure rate and the white balance of the camera properties. The blue hatched area shows the influence of shadows, which can falsify the colour value. Influencing factors of ambient light and shade must be taken into account in a later real-time correction. The current actual value can be compared with the known target value of the individual points in order to identify deviations and use them for further statistical evaluations. The Braille point described in the example illustrates one of many possible colour information from varying Braille points used in the calibration area (chapter 3.4) to adjust the camera.

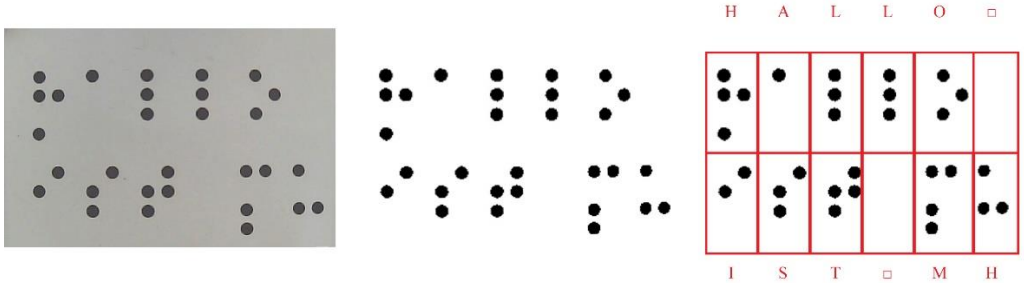


Fig. 4: Steps of image analysis : Left original colored image / middle: binarisation process after thresholding algorithm / right: pattern recognition (circle draw)

Figure 4 shows the original image (left), its binarized pattern by the Otsu's clustering-based image thresholding method (middle). The dots are grouped into the particular characters (right). This is achieved by measuring the distances between the dots and grouping them by vertical and horizontal segmentation. After identification of the individual groups and comparison with a Braille library, the characters are decoded.

3.2 Smart Code: Area recognition

Problem characterisation: In the following, we will demonstrate recognition algorithms that use a QR code as an example. The geometric form of the Finder pattern of the QR code is the same as in our smart code. By using a common static code such as the QR code, we want to provide an easier understanding of the following steps. Area recognition in a smart code involves the recognition of position markers (finder pattern) and the retrieval of data (static) and sensor (dynamic) areas.

Typically, the image of the camera captured the code is shaded, partial, or blurred. The code itself as well as the distance of single dots to each other or their patterns are distorted in both directions. All these erroneous parameters must be corrected before the code can be decrypted. High demands are placed on these corrections, especially with regard to robustness and reliability. Various suitable methods and algorithms have to be applied for the individual steps of error correction: e.g. the Circle Hough Transformation (geometric correction) described above or the Canny Algorithm (background suppression, edge detection and edge correction). These and other methods are available in program libraries and must be integrated operatively during programming.

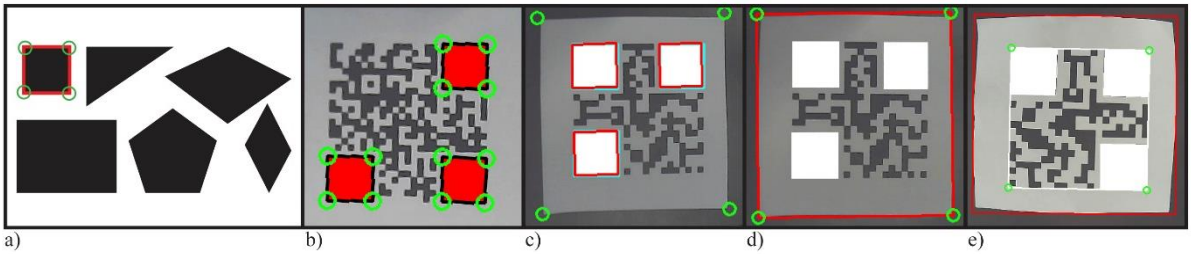


Fig. 5 Position marker recognition by geometric shaping

By recognition of geometric patterns, the individual position markers can be identified and therefore the data area as well as the sensor areas of a dynamic smart code can be located. This is particularly demonstrated in the prototype of the smart code in Figure 6. Based on the relative distances of the position markers distortions of the code can be balanced out. If the code is rotated, this will also be corrected. The square bars below and above the position markers identify the smart code as a code differing from a similar random pattern. In addition, the dynamic range for sensitive colours is identified.

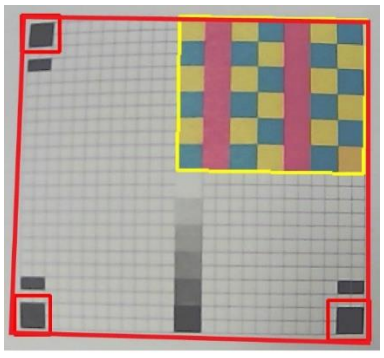


Fig. 6 Smart code prototype

3.3 Dynamic Data: Colour recognition

The following section deals with the evaluation and identification of the dynamic area (printable sensors) of the smart code. The respective Smart Materials are filled into the printer's cartridges instead of cyan, magenta and yellow. These three cartridges will contain a photochrome ink (photosensitive), pH-chrome ink (acid sensitive) and hydrochrome ink (humidity/ water sensitive). The black ink cartridge contains regular ink to prints all the static information of the code, including the finder pattern. The colour-

separated areas (by delimiting the colour ranges), such as the yellow squares, three magenta lines and the cyan squares represent the smart materials and their cartridges (Fig.9).

By detecting the dynamic area of the smart code, the colour values - of a specific position - inside of the dynamic code can be analysed. The status (degree of the contamination) of the individual smart material can be interpreted by means of the colour information. The aim is to identify the current colour values of each smart material and compare them with the measured colour reference before. This way, it is possible to refer the specific colour values with the dimension of the contamination. In this context, as well all RGB colour information of the background can also be used for colour correction of the dynamic code in order to get finally the true colour information. However, the colours in figure 7 show significant deviations, due to the camera detection. This indicates the necessity of a calibration of the colour values of the camera. This is planned to be realised in one of the next project steps by application of a neural network.

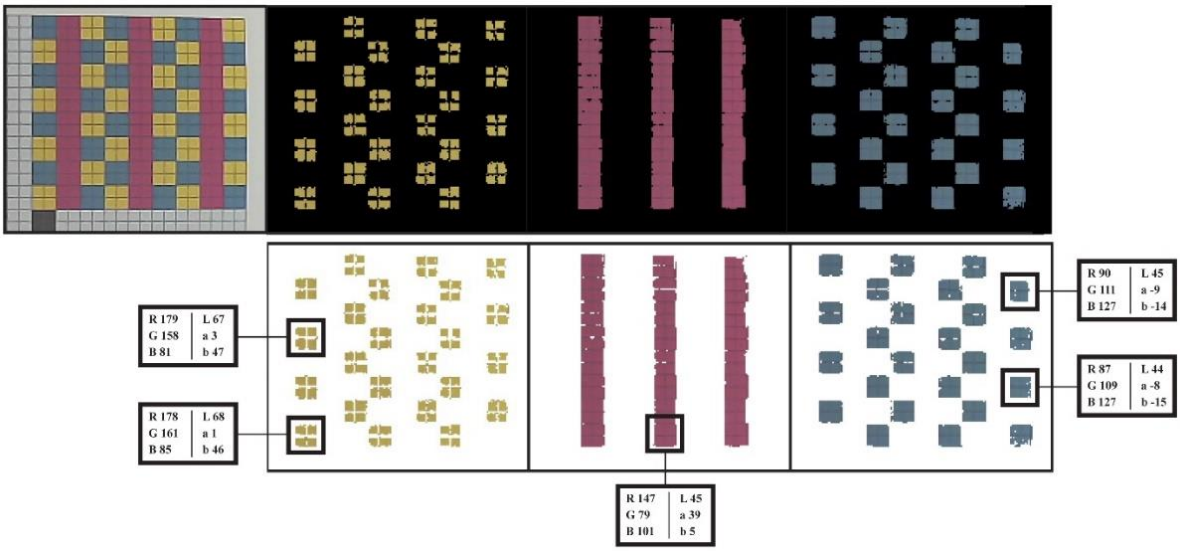


Fig. 7 Color detection and separation

The photochromic surface in figure 8 can be analysed by measurement of the RGB and CIEL*a*b values or alternative by RGB greyscale and L* values. Therefore it is possible to store information about critical deviations in a range from 0 (white) to 255 (black) in RGB or 0 (black) to 100 (white) in L*. A smart device can identify the RGB or CIEL*a*b* values of the different colour shades (smart dots) and compare them with the initial RGB or CIEL*a*b* values and set the colour difference (ΔE) into correlation with the irradiated quantity of light (photons) of the contamination.

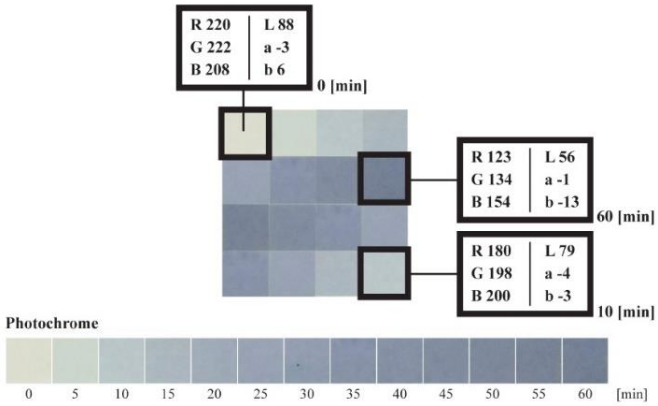


Fig. 8 Smart Material: Photochromic Ink

Table 1 shows the measurement data of the photochromic surfaces. A non-contaminated surface can be defined with an (R, G, B) tuple of (220, 221, 208) and a maximum exposed surface (60 minutes or longer UV exposition) can be described as a tuple of (121, 129, 150); 10 minutes exposition RGB is (174, 182, 188). The data of Table 2 allow deducing on the UV contamination of a product. As well, areas can be grouped in order to classify the quality of a product, e.g. good, moderate, and poor. Another application may be the indication of the entire UV intensity during the crosslinking process of UV printing ink or UV varnish or at crosslinking of UV adhesives.

Min of UV exposure	0	5	10	15	20	25	30	35	40	45	50	55	60
R	220	207	174	189	163	156	149	140	137	131	128	126	121
G	221	214	182	197	172	163	157	148	145	139	136	133	129
B	208	205	188	199	183	181	176	168	166	157	155	150	150
L*	88	85	79	74	70	68	65	61	60	58	57	54	54
a*	-2	-5	-4	-3	-2	-1	-2	-1	-1	-1	-1	-1	-1
b*	7	4	-2	-5	-8	-11	-12	-13	-11	-12	-12	-11	-13
ΔE	12.8	6.1	5.9	5.7	22.3	26.1	2	4.5	1	1.7	3.3	1.4	

Table 2 Measurement data of photochromic surfaces when UV exposed

3.4 Calibration

In the following, the calibration of the camera is discussed. The aim was to control the exposure and the white balance with a kind of colour checker function. The defined grey tones in figure 9 were integrated into the smart code and can be used for the calibration process. By capturing the smart code by using a code reader, the small areas within the grey areas (5 x 5 pixels) are captured in real time. Thus, the grey values and also the black values (Braille dots) in the data area are used to realize a correct exposure and white balance.

This was realized as described below:

First, the grey fields and their respective RGB colour values were measured and an average value determined (actual state). Next, the respective RGB averages were compared with those of the reference table (target state) and the respective deviation ΔE was calculated. The deviation ΔE was used as a colour check function to control the OpenCV exposure and white balance functions. The process was repeated as long until an acceptable exposure and white balance was achieved.

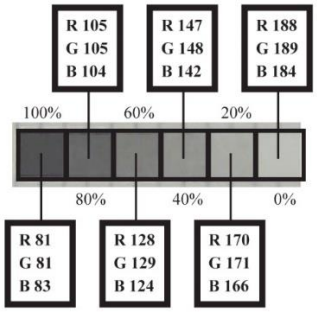


Fig. 5 Correct exposure and white balance using the colour checker function

Distortion manifestations such as barrel or fish eye effects have also to be corrected. The distortion matrix can be determined and corrected by using the asymmetric circle pattern resulting from the data range (Braille dots). Different lighting conditions - under the influence of light and shadow - have a significant effect to build a robust colour recognition system while doing colour interpretation.

4. Conclusions

It is possible to generate, print and read out a dot matrix code containing dynamic fields of sensitive inks with the camera of a smartphone. Furthermore, the colour information of the dynamic areas can be interpreted and correlated with corresponding contaminations. In laboratory scale, individual steps were shown to develop a smart code and to read it with a smartphone camera. To what extent the developed methods prove to be robust and to what extent the degree of several simultaneous contamination with the methods can be referenced and reproducibly determined must be demonstrated in comprehensive laboratory tests, whereby the developed process procedures must also be evaluated and readjusted. As soon as this will be verified, the data flow and the server operations can be designed and implemented in the IoT.

Here, some key steps are shown to develop a multifunctional sensor that works without its own power supply. The reading of the sensor information is possible by means of consumer devices and the interpretation of the information does a server within Internet of Things. Today, low availability of suitable, irreversible, sensitive dyes, which are inkjet print-able is problematically, yet. However, it should be supposed that this bottleneck quickly will be overcome as soon as these multifunctional, current less sensors have been developed beyond the laboratory scale and a market potential is identified.

References

Ashton, K. (2009). That 'Internet of Things' Thing; In the real world, things matter more than ideas. Hauppauge, New York, USA: RFID JOURNAL LLC. Retrieved from: www.rfidjournal.com /articles/view?4986. 06.03.2019

Bilgin, M. and Backhaus, J. (2017, a). Smart Packages by Means of Intelligent Codes. Advances in Printing and Media Technology. Advances Vol.44. Vol. XLIV(IV). Fribourg, Swiss: iarigai. ISBN 978-3-9870704-1-9. ISSN 2409-4021. p.89-96

Bilgin, M. and Backhaus, J. (2018). Development of a unidirectional switchable Photochromic Ink for Smart Packaging. Advances in Printing and Media Technology. Advances Vol.45 (2018). (Online) Vol. XLV(V), iarigai, Warsaw, Poland. ISBN 978-3-948039-00-4. ISSN 2409-4021. P.55-63

Hough V, Paul C. (1962). United States Patent 3069654, Application Number: US1771560A, Publication Date: 12/18/1962

International Standard ISO/IEC 10646 – Unicode. (2010). ISO 11548-1:2001. Communication aids for blind persons – identifiers, names and assignation to coded character sets for 8-dot Braille characters – Part 1: General guidelines for Braille identifiers and shift marks. Retrieved from: www. unicode.org/L2/L2010/10038-fcd10646-main.pdf. 06.03.2019.

OpenCV. (2019). About OpenCV. Retrieved from: www.opencv.org/about.html. 06.03.2019.

Pascolini D. and Mariotti SP. (2012). Global estimates of visual impairment. British Journal of Ophthalmology (BJO). London, UK: BMJ Publishing Group. May, 96(5):614-8. doi: 10.1136/bjophthalmol-2011-300539. p.1-5.

Yuen, H. K.; Princen, John; Illingworth, John; Kittler, Josef. (1990). Volume 8, Issue 1, February, Amsterdam, Netherlands: Image Vision Computer. DOI:10.1016/0262-8856(90)90059-E. p.71-77

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4467-4481
https://publishoa.com
ISSN: 1309-3452

Visually Impaired Aid using Computer Vision to read the obstacles

Pragati Chandankhede

Arun Kumar

[1] PhD scholar, Department of CSE, Sir Padampat Singhania University, Udaipur

[2] Professor, Department of CSE , Sir Padampat Singhania University, Udaipur

Received 2022 March 15; **Revised** 2022 April 20; **Accepted** 2022 May 10.

Abstract:

The world for normal human being is far different than visually impaired, due to either lack of vision or no vision. The difficulties in their daily routines can be minimized with help of technological support which is usually aids that can be used for travelling. Computer vision a field of artificial intelligence provides the assistance for helping impaired. The assembly of the device is made as handy as of user is using mobile and the architecture of YOLO (You Only Look Once) is used for accurate object detection. The feature detection of YOLO is more appropriate in real time. The object or the obstacle from which user can collide or the pedestrian localization is indicated to user with help of speaker which make system valuable. The compact size, Raspberry Pi 4 B 8 GB is used or processing, which has proved accuracy of 98% in real time scene.

Background:

The structure of designing a system for visually impaired to classify the aspect of the visual scene which represent the most important features for navigation and object identification (presence of the objects and their position in space). The auditory system, which is capable of combining information by classes of clues, plays a crucial role for the navigation. The thought of Computer vision is to acknowledge and interpret pictures an equivalent approach humans do, distinctive them, classify them, and type them supported their characteristic traits, like size, color, then forth. Images play a big role in human perception. However, in contrast to humans, computers or machines rework a picture into digital kind and perform some method thereon to urge some substantive data out of it.

Objectives:

Objectives identified were

1. Designing the compact size device that works on real time for visually impaired.

2.Cost effective system for understanding scene.

Keywords: travelling Aid, sensor based technology, navigation system, computer Vision, object detection, deep learning, pattern recognition, convolution neural network

1. Introduction

Visual impairment can be defined as the situation where eitherperson either not having ability to see or his vision has weakened to large extent. According to World Blind Union, there exist 314 million of blind and partially sighted people in the world. Out of which 45 millions are declared as completely blind. Reason for the blindness varies, but majorly exist due to neurology parameters or physiological. Thus those people with Low vision or blindness impacts their day to day life activities. Not only daily only chores but social communication also trims down due to reduced mobility. The majority of the primary causes of vision impairment includes cataract and under corrected refractive error, are subject matter to the epidemiological transition. This can be defined by changing pattern in various age group that impacts directly individual with societal costs. The blinding ophthalmic circumstances that are marked are called

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4467-4481
https://publishoa.com
ISSN: 1309-3452

epidemiological transition that includes age-related cataract, or when part of retina called macula is damaged and complication to diabetes patient that influence eyes[1].

With an blindness estimation of about 76 million by 2020, there arose a universal need that can help against circumstances . Major organization like International Agency for Prevention of Blindness and WHO have initiated progam "Vision 2020: The Right to Sight which take care of various national eye care plans. In 2013, the World Health Asssembly (WHA) launched a new plan, Towards universal eye health: a global action plan 2014–2019 (GAP). It set a global target: to achieve by 2019 a 25% reduction from the baseline of 2010 in prevalence of "avoidable" visual impairment, defined as the aggregated crude prevalence of cataract and undercorrected refractive error. These effort are acquired to since Vision impairment and blindness are directly associated with reduced economic, educational, and employment opportunities that affect quality of life. Being developing country like India, which get improved in progress on the scale of not only socioeconomic development but the way life seems to be satisfied with expectancies. As per National Programme for Control of Blindness & Visual Impair-ment(NPCBVI),Directorate General of Health Services, Ministry of Health & Family Welfare, Government of India, there are 1,14,33,232 blind people registered in India till 2022(data obtained from https://npcbvi.gov.in/). It was found that hereditary diseases, Diabetic retinopathy, glaucoma, squint, keratoplasty, retinopathy with prematurity, congenital ptosis and cataract were the prominent causes in India for impairment.With immense greater part of vision impairment and blindness were caused through diabetic retinopathy, cataract, and glaucoma can be evaded with early detection and timely intervention.

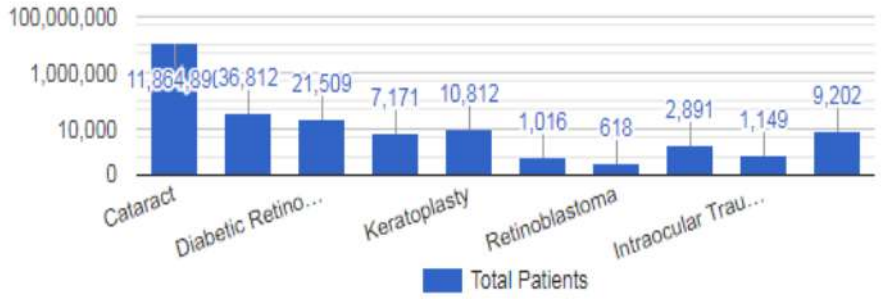


Figure 1: Total Registered Patient in India

2. Literature Review

2.1 Navigation System

In twentieth era for normal being, communication via social networking website are popular, the basic requiring like "going for walk" seems difficult for the visually person. The concept of orientation and way finding depends on how we avoid obstacles to do best mobility. Normally, individuals depend on vision to know their own position and which direction to move in the environment, they can recognize numerous elements in their surroundings, as well as their distribution and relative location. But can we imagine a mobility being carried out without vision? Vision less human being or low vision being finds difficulty and his performance slow down for basic tasks of finding his/her own stuff in his/her room. But with consistent efforts, blind person incorporate awareness from remaining sensory system, memories, verbal imagery to make life easy. The device or machine that can help the person to reach destination point is said as navigation system. Recent years had shown extraordinary improvement in tools that can help mobility.

Ability to travel and mobility of blind depends on four main factors, obstacle detection, and environment mapping, and navigation, relative location [5]. Over the decades, researcher were working on prototype of obstacle detection system, huge success in this was due to sonar and radar system. In both the phenomena the basic is determining target distance from the user and alerting user. Broadly travelling aids are categorized as Type I and Type II, both having unique features in them.Some prominent system of Type I and Type II are illustrated below.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4467-4481
https://publishoa.com
ISSN: 1309-3452

2.2 Navigation system: Type I Electronic Travelling Aid

1. The Russell Pathsounder

The Russell Pathsounder invented in year 1966, by the author Lindsay Russell. The device has proved successful with exceptional applications, as it facilitate a man with low vision to retain his job as a floor supervisor in a sheltered workshop. The device has enabled a small boy who was neurologically impaired completely dependent on a support cane, to develop outdoor orientation and mobility skills. The phenomena underlying was, the use of ultrasonic waves into space at the rate of 15 pulses per second. The output here is either vibratory (tactile) and auditory[Li Kun , (2015)]. Device will haveno output if if waves do not collide with any reflecting surface which measures till length 79-182 cm. Developing a creative approach to the rehabilitation of visually impaired clients withcomplex needs. While the limitation was use of single output acoustic waves for object preview.

Figure2: The Russell Pathsounder

2. Ultrasonic Cone:

Invented in year 1972 by the researcher Geoff Mowat uses circular transducer, forward range to 4.02 meters. Also provided a switch to shorter distance range of about 1 meter. The acoustic waves creates divergence cone that cover up the size of body of human. However, there are certain problems with the device, which may have false readings under cold, rain, heavy snow weather conditions. The detecting regions by the cone are depicted in below figure.

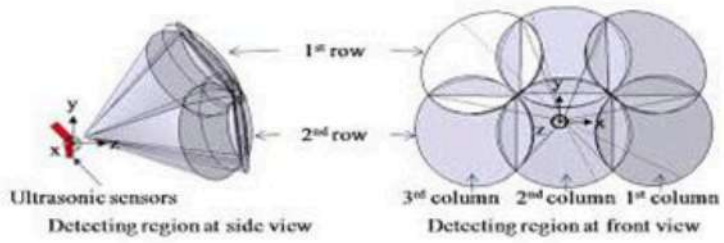


Figure 3: Ultrasonic Cone

3. Polaron:

The Nurion Industries brought out Polaron in 1980s. It can be applied on hand or chest. This devices make the use of laser cane that ranges selection of 1.22, 2.44, and 4.88 meters (4, 8, 16 feet), and an obstacles that are ahead are guided by vibrotacile or audible signals[3]. Problem with device was it cannot give any indication of the actual distance of the obstacle.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4467-4481
https://publishoa.com
ISSN: 1309-3452

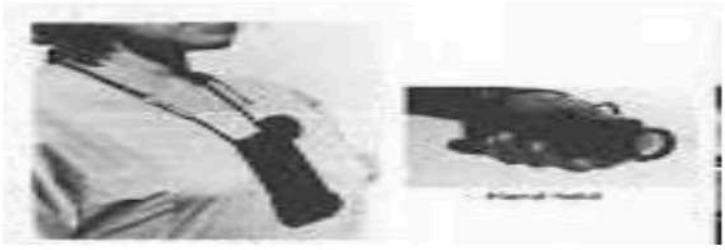


Figure 4: Polaron

4. WalkMate:

Developed in 1993, it emits U shaped beam of 0.7 meter wide and 1.82 meter tall. This u shaped beam maps the obstacles ahead. The shape was designed as 'u' which tries to acquire major objects ahead of it. But this model was not too effective in real life case and hence it's not available.

Figure 5: WalkMate

5. Miniguide US:

MiniGuide US is the only Type-I products still for sale these days. This product started as least from year 2004. It Provides a range of 7.92 meters, and the price was $545 online. Limitation of the system was as plastic case is used, chances of breakage. Being small in size, lack the coordination for control manipulation and accurate positioning.

Figure 6:MiniGuide US

6. vOICE:

Invented by MalikaAuvray, Sylvain Hanneton, J Kevin O'Regan, the technique is implemented for purely blind person, by the art of understanding images ahead through the process of rendering digital senses that lead to synthetic vision by true visual sensations through cross modal sensory integration.Such a technique need training.The phenomena is based on cognition.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4467-4481
https://publishoa.com
ISSN: 1309-3452

Figure 7: vOICE

2.3 Navigation system: Type II Electronic Travelling Aid

This type of ETA primarily focuses on smart sticks, white can and smart dogs. Use of long cane tries to map the environment for about 1.5m with wind resistant capacity and its design in fashion so that it should be visible to pedestrians. Even though use of cane is simplest but it's have disadvantage that it can fetch enough information about big obstacle ahead. With the emergence of electronics tool after world war second, this simple cane is transformed to smart cane. The detail about smart cane is as shown below.

Type II Electronic Travelling Aid published year along with description and limitation of mentioned system are as follows.

1. White cane:

Invented in year 2001 by C. Wong, D. Wee, I. Murray and T. Dias has proposed design centres around the use of a micro controller to calculate distance measurements of ultrasonic signals sent and received by two sets of transducers fixed onto the white cane. The limitation of system are 1.The ability of a user to detect a dip is based on the user's sensitivity to ground levels 2.When travelling in crowded places the cane should be kept close to the body to avoid tripping other people.

2. Smart cane:

Invented by Whitney Huang, Hunter McNamara in the year 2014 uses Ardumoto to control Pulse Width Modulation. The Ardumoto can control vibrating motors in many ways through an analog input for the motor speeds, on and off features, and direction features. the device utilizes the information from an accelerometer to detect the orientation and motion of the hand. The limitation of the system was Accuracy of the distance reading is affected when a moving ultrasound source tries to detect a stationary object. The primary difficulty with creating a controlled experiment is maintaining a constant sweep rate with the cane.

Figure 8: Smart cane

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4467-4481
https://publishoa.com
ISSN: 1309-3452

3. Intelligent walking stick:

Invented by Nadia nowshinsakibshadmanSaha joy in the year 2017 has four ultrasonic sensor stick have sensor a and b implemented on the front side sensor c on the left side and senor d on the right side all four sensors detect obstacles and send the distance to an Arduino module then Arduino sends the correct distance to the mobile app through Bluetooth module then through the app user can hear the distance in their headphone.The system can be more effective with image processing.

3. Contextual Information needed by Visually Impaired

The way normal human being needs the information to take actions, the same way Contextual information is wanted by visually impaired innew, unfamiliar spaces outside their home. The information usually is the sound that they can follow to find the objects. Or the information is touch senses that they can feel via hand. Specifically in known environment, they can even tell the object they are touching. Sometime this information is tag along the path that came across them, by counting the number of steps they have till followed, that can lead them to valid destination. But the information needed by visually impaired is not restricted to above mentioned, it conforms from going to market place for purchasing basic needs, and crossing road. For which they have to rely on someone. Understanding the public services such as transport is also an important detail.

For visually impaired, out of many sectors of application, the education is the field which was considered as important and technological support has even strengthen the roots of this sectors. Braille was foremost effective technique for learning numbers and alphabets [4]. Following figures are demonstration of the same. The Braille is read by index figure rolling on the each line, as normal human being start writing from left to write its also read in same way. With Braille we can make aware to blind about punctuation or spelling and even the direction indication at public places. Efforts in such a direction were taken by IIT Delhi. By deploying Dot Book which is a Braille display. This Dot Book can give illustration for digital content or routines that can provide comfort in working environment.

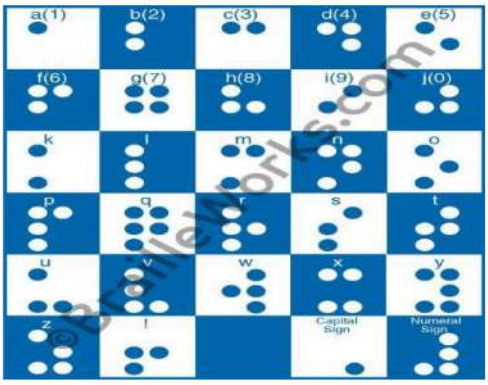


Figure 9: Braille work flow for learning alphabets and letter

Figure 10: DotBook

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4467-4481
https://publishoa.com
ISSN: 1309-3452

4. Present Navigation facility for Visually impaired at public places

This is the a way to be ready for written info that too in Braille, though once true square measure of real time scene like getting to malls, museums, there's so want of Navigation facility. The additional challenge is because of enough complexness through within house and lack of coaching in mobility. one amongst the incapacity startup of Asian country, AssisTech Foundation fabricated smart cane helps out these individuals to nice extent. This is often associate ultrasonic travel device which will be connected to a customary white cane for detection of knee on top of obstacles moreover as non-contact detection, it's sixty thousand user in Asian country[9]. Laws mandating access to public places by the Constitution of Asian. country states the security provision that require to be followed Sections 44-46 of the Persons with Disabilities Act, 1995 emphasize the importance of providing non-discriminatory access by removing all physical barriers. The access to public places within the following ways in which ought to follow following points

A. Apropriately fixing buses, airplanes, train compartments and vessels.

B.To form them accessible to persons with disabilities through chair user

C. putting in modality feedback in traffic signals for the advantage of the visually impaired.

D. creating necessary curb cuts and slopes in pavements for chair users.

E. Engraving the surface of equid crossings for the visually impaired.

F. Engraving the sides of railway platforms for the advantage of the visually impaired.

The malls have the fix structures, in such a case the "flow" info (example, voice message is passed on to user " you're on ground level, initially floor you'll obtain occasional, you're on second floor you'll obtain suits and dress, T shirt and Jeans")that is on the market is exploited to supply effective navigation support to the guests. however such location primarily based info is difficult as there'll be following with facilitate of either GPS, Bluetooth, or CCTV cameras. Such a system has not been nevertheless established.

5. Need of practicability for daily chores

All gratuitous obstructions should be removed, and every one access ways in which should be lit. Moreover, clear signposts, in conjunction with their Braille equivalents ought to be place up. Elevators should have clear Braille signs and modality feedback. The buttons of elevators should be accessible from a chair. Pictograms should be place up close to elevators and different necessary places like bogs. albeit a considerable portion of public places still stay inaccessible in Asian country, it's encouraging to notice that some samples of totally accessible buildings that square measure price emulating have appeared within the previous couple of years. The Old Delhi subway is probably the primary massive scale project in Asian country, a minimum of within the transport sector, that adheres to any or all accessibility standards and embraces an outsized array of best practices for the welfare of the disabled. The Chhatrapati Shivaji Maharaj Vastu Sangrahalaya, referred to as the patrician of Wales repository, in Mumbai is supplied with ramps, hydraulic lifts moreover as Braille aggregation for the advantage of the disabled.

6. Object Detection and Computer vision

To get insight for a entire image understanding, we must concentrate on categorizing different im ages, and make an attempt to exactly estimate locations of objects given in each image[2]. The process is said as object detection, which usually consists of different subtasks such as face detection, pedestrian detection. (Papert, Seymour A., 1966) proposed Ground Analysis and region analysis, the article was considered as a remark for stepping in direction for pattern recognition[8]. Over last 50 years, Computer vision has not been solved, and is still a extremely tough problem. We can define it as "a little that we humans do without thinking" but that is actually hard for computers to do or even to understand. Problems being hard because there is a enormous gap sandwiched between meaning and pixels. The

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4467-4481
https://publishoa.com
ISSN: 1309-3452

computer sees in a 200 × 200 RGB image is a set of 120, 000 values. The path from understanding these capability of visual system is incredibly proficient. Not only it recognizing fear and reacts to it immediately. And can we map of whether machine can do the same way as that of human can do? Answer was yes. Answer lies in roots of computer vision.Whether it's a computer or an animal, vision comes down by two components.

Foremost, sensing device that will capture details from a given image. For human, the eye will capture light coming through the iris and project it to the retina, where specialized cells will transmit information to the brain through neurons. Secondly, for a device with camera, captures images in a similar way and transmit pixels to the computer. In this part, cameras are better than humans as they can distinguish infrared, see far away or with more precision. Difficulty lies when device has to understand and process the information and extract meaning from it. The human brain resolves this in multiple steps in different regions of the brain. Computer vision is taking its step in direction to make it possible. Computer vision cannot work alone, with the technology of Artificial intelligence phenomena that manifests brain help out to lift results for image understanding.

Computer vision is a blended technology of pattern recognition and image processing. Image understanding is the focus of Computer vision. In order to make it work, we need to create models and extracts the data from the given images. Further, Image Processing is to transform them computationally. using an algorithm and sensor Computer vision has been motivated human visual system. But still it can't be as perfect as human. Along with gaining an image, analyzing image make computer vision more striking which involve steps 1) image creation, during this phase, image of object is captured and stored in computer; 2) image initialization/preprocessing, whereby quality of image is improved to enhance the image detail; 3) image segmentation, in which the object image is identified and separated from the background, 4) image measurement, where several significant features are quantized, and 5) image interpretation, where the extracted images are then interpreted.

A. MS COCO Dataset:

In order to enable a machine for recognizing object and then categorizing object, which include understanding patches, parts and features of given object, various dataset have been created. Dataset will also help into identifying each sort of a class from a broad collection of images.

MS COCO dataset (Common Objects in Context) is a large-scale object detection dataset that additionally includes segmentation, and captioning a part of dataset and is revealed by Microsoft [10]. So as to grasp visual scene, majority of Computer vision algorithmic rule used this dataset. The format of the dataset is mechanically taken by advanced neural network libraries that embody Tensorflow, PyTorch, OpenNN. There are eighty object classes aforesaid because the "COCO classes", that comprise "things" that individual instances is also merely labelled (person, car, chair, etc.) ninety one stuff classes, wherever "COCO stuff" includes materials and objects with no clear boundaries (sky, street, grass, etc.) that give important discourse data[Li Fei-Fei(2007)].

B. Object Detection Algorithm

To prepare algorithm seems being simple but second part to let them understand scene and react is quite difficult. It's because raw input to such learning system is always high dimensional entity. It may be solid object with 3Dimensional view. For such a case, computer vision and object recognition work together [2]. Image recognition is best handled by deep learning technique of AI, which can better serve visually impaired person.

Various object detection system come into existence, but the problem with many of them was, they depends on classifiers to be applied at multiple scale and location. The open source Neural Network, Darknet has presented many model, You only look once is one of them. Yolo resolvesthe problem by applying single network to full image. The network divides the image into regions and predicts the possibilities of object in an image with help of bounding box. The interesting part of Yolo v3 is its small feature map that packs lots of information within it. It can also be said as loss function, which has the capability of predicting same object of different size. The feature of IoU, Intersection over union value that notify that whether the object is present in said bounding box or not, it ranges 0 if not present and 1 if its perfectly able to predict. The purpose of loss function is to find most excellent IoU.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4467-4481
https://publishoa.com
ISSN: 1309-3452

C. Working of YOLO version 3

Image Detection method typically apply localizer and classifiers for correct detection of object. additionally detection method is increased by giving input at completely different scales and site[4]. The absolutely connected model of the YOLO algorithmic rule is completely different than previous approach. The algorithmic rule applies one neural network to the complete full image. The given image is split into regions and regions offers proposal of bounding box. With notion of presence or absence of object within the region the possibilities is calculated. There are 5 versions of Yolo. Out of that version 3 is employed for the project [3].

YOLO V3 is fifty three layer network trained on Imagenet. For the task of detection, fifty three a lot of layers ar stacked onto it, giving United States a 106 layer absolutely convolution underlying design for YOLO. the foremost salient feature of v3 is that it makes detections at 3 completely different scales. YOLO may be a absolutely convolutional network and its ultimate output is generated by applying a one x one kernel on a feature map. The form of the detection kernel is one x one x (B x (5 + C)). Here B is that the variety of bounding boxes a cell on the feature map will predict, "5" is for the four bounding box attributes and one object confidence, and C is that the variety of categories. In YOLO v3 trained on coconut palm, B = three and C = eighty, therefore the kernel size is one x one x 255. The feature map created by this kernel has identical height and dimension of the previous feature map, and has detection attributes on the depth as delineated higher than. YOLO v3, in total uses nine anchor boxes 3 for every scale[7].

D. Architecture of system

The figure 10, represent the architecture of the system, even though the process looks long , YOLO object detection algorithm completes the detection and prediction in few second, the real life result of YOLO are effective for many cases. The confidence that is calculated by YOLO will give us accurate object prediction. YOLO is trained on COCO dataset that has labels for 80 classes, despite of variation in angel, color and shape all 80 classes can be recognized.

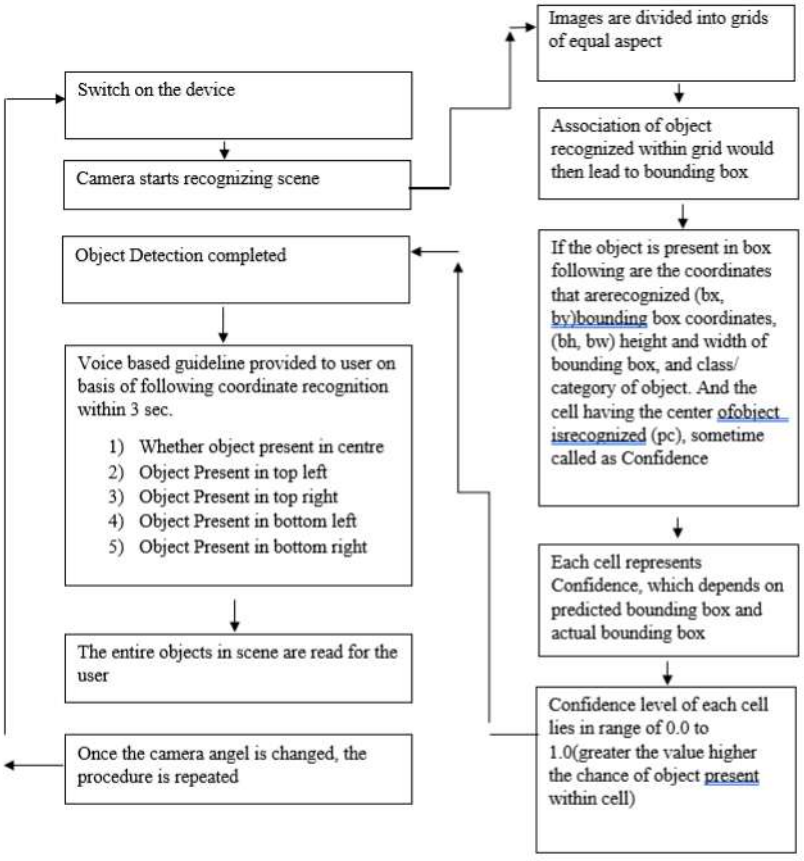


Figure 11: Architecture of system

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4467-4481
https://publishoa.com
ISSN: 1309-3452

7. Result and Discussion

A. Performing Object Detection with Predefined Weights and configuration of YOLO.

In today's world, RAM on a machine is reasonable and is on the market in lots. You wish many GBs of RAM to run a brilliant complicated supervised machine learning downside – it is yours for a bit investment / rent. On the opposite hand, access to GPUs isn't that low cost. you wish access to hundred GB VRAM on GPUs – it won't be undemanding and would involve important prices. we tend to attempt to solve complicated reality issues on areas like image and voice recognition. Once you have got some hidden layers in your model, adding another layer of hidden layer would want huge resources. A pre-trained model could be a model created by some one else to resolve an analogous downside. rather than building a model from scratch to resolve an analogous downside, you utilize the model trained on alternative downside as a place to begin.Pre-trained networks demonstrate a powerful ability to generalize to image.

Deep learning techniques area unit achieving progressive results for object detection, like on customary benchmark datasets and in computer vision competitions. Notable is that the "You only Look Once," or YOLO, family of Convolutional Neural Networks that deliver the goods close to progressive results with one end-to-end model which will perform object detection in period of time.[6]

Traditional systems repurpose classifiers to perform detection. Basically, to sight any object, the system takes a classifier for that object then classifies its presence at varied locations within the image. Alternative systems generate potential bounding boxes in picture mistreatment region proposal strategies so run a classifier on these potential boxes. This leads to a rather economical technique. when classification, post-processing is employed to refine the bounding boxes, eliminate duplicate detection, etc. because of these complexities, the system becomes slow and exhausting to optimize as a result of every part needs to be trained on an individual basis.

1. The primary step is to transfer the pre-trained model weights.

2. Offer the configuration and weight files for the model and load the network.

3. Load names of classes/labels

4. calculate the network response for blob

5. Loop over every of the layer outputs

6. Separate out weak predictions by guaranteeing the detected chance is larger than the minimum probability (confidence>0.4)

7. Convert detected image with speech as a voice output for direction and discovering obstacle or object ahead. Drive top left corner of bounding box.

8. Apply non-maxima suppression for overlapping boxes.

9. Guarantee a minimum of one detection.

The RaspberryPi 4 B 8 GB is used for processing object detection algorithm since the current era depends on compact size architect devices. This gives a chance to user to create something innovative. The Raspberry Pi is computer of shape as small as debit card with following key feature, 2 × USB 2.0 Ports 2 × USB 3.0 Ports, 2.4 GHz and 5.0 GHz IEEE 802.11b/g/n/ac wireless LAN, BLE Gigabit Ethernet, Bluetooth 5.0. And the operationg power is 5V 3A DC via GPIO Header 5V 3A DC via USB Type-C Connector Power Over Ethernet (PoE)–Enabled.use of c5-megapixel OV5647 Raspberry Pi Camera that even works in low light is used along Pi.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4467-4481
https://publishoa.com
ISSN: 1309-3452

Figure: 11 The Raspberry Pi with Camera Module and Connection for Object detection

To access the Raspberry Pi command-line interface, PuTTY is used. It uses SSH (secure shell) to open a terminal window on your computer, which you can use to send commands to the Raspberry Pi and receive data from it. After inserting Yolo v3 setup in Raspberry Pi, whole configuration is made customize by three dimensional printing. The overall system ready for use seem to be, as shown in figure 11.

If visually impaired wish to connect it to Television for recognizing object in a living room, where every time he need not to on or off the system, that's possible with the approach. In all the cases object detected by the system will be informed to the user via headphone, so he need to apply them. Advantage of this approach is small objects which are located inhouse like watch, specs can be found by the person by simply moving the device camera around the room. And device will guide the user that specs are in bottom right corner.

Figure 12: Handy Device for Visually Impaired

4 x 4 x 1 inches is the totality size of device shown in above figure, which is same as medium size mobile and hence such a handy device can be carried and used whenever user who is visually impaired needs and environment is new to them. Following are the real life result obtained by the system.

Figure 13: Object Detection by the system

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4467-4481
https://publishoa.com
ISSN: 1309-3452

Figure 14: Visually impaired holding Device for detecting object ahead.

8. Conclusion and Future work

Various Travelling Aid have made mobility possible for visually impaired, but still there is lot of scope for up gradation, the work presented in paper is one step in the direction. By many surveys it was proved that visually impaired can hear better, sense better and can make themselves in scenario where they are safe. Taking into this consideration, the device aims of "telling" whatever it sees through camera module of Raspberry Pi 4 B 8 GB . The System has proved to be effective in real life scene with precision of 0.496 and recall of 0.623 with intersection over union for the range of 0.50 to 0.95. The key parameter of YOLO is, it understand contextual information about scene within minimum time. The location of object detected is alerted to the user with voice guideline providing either of five coordinate positions.The key features of the system is compact size device that can read obstacles can make visually impaired more independent that they can rely on. Future work involves optimizing the algorithm for the optical character recognition and ready Braille characters applied at public places. Future work involves Image magnification with computer vision algorithm can benefit many person who are not completely blind.

References

[1] Amira B, Ramadan T, "Face Detection and Recognition Using OpenCV", Journal of Soft Computing and Data Mining,VOL.2 NO. 1, 86-97, October 2021. https://www.researchgate.net/publication/355886757_Face_Detection_and_Recognition_Using_OpenCV.

[2] Dixit R, kushwahR, "Handwritten Digit Recognition using Machine and Deep Learning Algorithms", arXiv, 2106.12614, June 2021. https://arxiv.org/abs/2106.12614.

[3] FarhadiA. ,Redmon J. "YOLOv3: An Incremental Improvement", arXiv,1804.02767, Jan 2018. https://arxiv.org/abs/1804.02767.

[4] Kavalgeri S, "E - Braille: A Study Aid For Visual Impaired", International, Journal of Research in Engineering, Science and Management, volume 2, Issue 2, April 2019. https://www.researchgate.net/publication/332571033_E_-_Braille_A_Study_Aid_For_Visual_Impaired.

[5] "An Industry Landscape Study Electronic Travel Aids for Blind Guidance", Sutardja Center for Entre-preneurship & Technology Technical Report,1-10.Dec 2015.

[6] Li Fei-Fei, Fergus R. , Pietro P., "Learning generative visual models from few training exam-ples:An incremental Bayesian approach tested on 101 object categories", Journal of Computer Vision and Image Understanding, 106 59–70,April 2007. https://ieeexplore.ieee.org/document/1384978.

[7] Pathak A. , Pandey M., Rautaray S, "Application of Deep Learning for Object Detection", Inter-national Conference on Computational Intelligence and Data Science , 132 ,1706–1717, 2018. https://www.sciencedirect.com/science/article/pii/S1877050918308767.

[8] Rupert R, "Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight" An analysis for the Global Burden of Disease Study, Blindness and Vision Impairment Collaborators, volume 9 issue 2 .2021. https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(20)30489-7/fulltext.

[9] Real S, Araujo A "Navigation Systems for the Blind and Visually Impaired: Past Work, Challenges, and Open Problems", International Journal of Sensors, 2019,3404.", International Journal of Research in Engineering, Science and Management, 2581-5792. June 2019. https://www.mdpi.com/1424-8220/19/15/3404.

[10] Tsung-Yi Lin, Michael Maire et,al, "Microsoft COCO: Common Objects in Context", ar-Xiv,1405.0312, 2015. https://arxiv.org/abs/1405.0312.

Pragati Chandankhede , PhD scholar at Sir Padampat Singhania University and presently working as Assistant Professor in KCCEMSR Thane. She holds master's degree in Computer Science and Engineering from G.H. Raisoni College of Engineering and Bachelor's degree in Information Technology from Amravati University. She has interest in the streams of Artificial Intelligence, Computer vision and Cloud Computing. She has guided various BE projects. Has Presented papers at various International conference and published the research article in Journals

Prof Arun Kumar is presently working as Professor in the Department of Computer Science and Engineering at Sir Padampat Singhania University, Udaipur, Rajasthan. He is also shouldering the responsibility of Dean of School of Engineering at SPSU. He holds a Bachelor's degree in Applied Electronics and Instrumentation Engineering from NIT Rourkela, a master's degree in Computer Science and Engineering from University of Madras and a Doctoral degree in the area of Computer Vision. He has interest in the development of application in the area of Data Science, Recommender Systems, and Fake News Analysis. He holds two granted patents from Govt of India. He is a registered PhD guide with SPSU and has four research scholars who have already graduated from the department of Computer Science and Engineering

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Mi in nulla posuere sollicitudin aliquam. Egestas diam in arcu cursus. Tincidunt arcu non sodales neque. Id neque aliquam vestibulum morbi. Donec enim diam vulputate ut pharetra sit amet aliquam id. Enim sed faucibus turpis in eu mi bibendum neque egestas. Sed enim ut sem viverra. Donec ultrices tincidunt arcu non. Varius sit amet mattis vulputate enim nulla aliquet porttitor. Ultrices dui sapien eget mi proin sed libero enim. Sem viverra aliquet eget sit. Malesuada nunc vel risus commodo viverra maecenas accumsan lacus vel.

Quis risus sed vulputate odio ut enim. Laoreet suspendisse interdum consectetur libero id faucibus nisl. Egestas maecenas pharetra convallis posuere morbi. Vitae suscipit tellus mauris a diam maecenas. Sit amet cursus sit amet. Dui nunc mattis enim ut tellus. Amet nulla facilisi morbi tempus iaculis. A iaculis at erat pellentesque adipiscing commodo elit at imperdiet. Pulvinar mattis nunc sed blandit libero volutpat sed. Tincidunt ornare massa eget egestas purus viverra accumsan in nisl. Fermentum odio eu feugiat pretium. Tellus mauris a diam maecenas. Tincidunt lobortis feugiat vivamus at. Tincidunt tortor aliquam nulla facilisi cras. Enim neque volutpat ac tincidunt vitae. Amet massa vitae tortor condimentum. Ut tortor pretium viverra suspendisse potenti nullam ac tortor. Convallis aenean et tortor at.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4467-4481
https://publishoa.com
ISSN: 1309-3452

16. Methods

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Orci a scelerisque purus semper eget duis at tellus at. Quisque egestas diam in arcu cursus. Pulvinar mattis nunc sed blandit. Tempus iaculis urna id volutpat lacus laoreet non curabitur. Morbi tincidunt ornare massa eget egestas purus viverra accumsan in. Vehicula ipsum a arcu cursus. Sapien et ligula ullamcorper malesuada proin. Ut diam quam nulla porttitor. Tincidunt dui ut ornare lectus sit. Neque ornare aenean euismod elementum nisi quis eleifend. Mus mauris vitae ultricies leo integer. In nulla posuere sollicitudin aliquam ultrices. Eget duis at tellus at urna condimentum mattis. Tellus molestie nunc non blandit. Quam quisque id diam vel quam elementum pulvinar. Integer quis auctor elit sed vulputate mi. Pellentesque elit eget gravida cum sociis natoque penatibus et. Aliquet risus feugiat in ante. Commodo ullamcorper a lacus vestibulum sed.

Congue nisi vitae suscipit tellus mauris a diam maecenas. Aliquet nec ullamcorper sit amet risus. Pulvinar sapien et ligula ullamcorper malesuada proin libero nunc consequat. Non consectetur a erat nam at lectus urna duis convallis. Purus viverra accumsan in nisl nisi scelerisque eu. Netus et malesuada fames ac turpis egestas maecenas pharetra convallis. Sed turpis tincidunt id aliquet. Et malesuada fames ac turpis egestas sed tempus urna et. In dictum non consectetur a erat nam at. Nulla aliquet porttitor lacus luctus accumsan tortor posuere. Nunc consequat interdum varius sit amet mattis vulputate enim nulla. Cras tincidunt lobortis feugiat vivamus. Venenatis a condimentum vitae sapien pellentesque habitant morbi. Suscipit adipiscing bibendum est ultricies integer. Et ultrices neque ornare aenean. Ut porttitor leo a diam sollicitudin tempor id eu. Lorem ipsum dolor sit amet consectetur adipiscing elit. Morbi tincidunt ornare massa eget egestas purus viverra accumsan in. Sit amet consectetur adipiscing elit duis tristique.

Ipsum dolor sit amet consectetur adipiscing. Arcu felis bibendum ut tristique. Lectus sit amet est placerat in egestas. In massa tempor nec feugiat nisl pretium. Vel pharetra vel turpis nunc eget lorem dolor. Ornare aenean euismod elementum nisi quis eleifend quam. Tellus id interdum velit laoreet id donec. Eget arcu dictum varius duis at consectetur lorem donec massa. Amet facilisis magna etiam tempor orci eu lobortis. Consectetur adipiscing elit duis tristique sollicitudin. Pellentesque dignissim enim sit amet venenatis urna cursus eget.

Pellentesque adipiscing commodo elit at imperdiet. Lectus proin nibh nisl condimentum id venenatis. Dignissim diam quis enim lobortis scelerisque fermentum dui faucibus in. Volutpat diam ut venenatis tellus. Vehicula ipsum a arcu cursus vitae. Volutpat maecenas volutpat blandit aliquam etiam. Sed id semper risus in. Eget nulla facilisi etiam dignissim diam quis enim lobortis scelerisque. Tellus in hac habitasse platea dictumst. Non enim praesent elementum facilisis leo. A cras semper auctor neque vitae tempus quam pellentesque. Dolor magna eget est lorem ipsum dolor sit amet consectetur.

Neque laoreet suspendisse interdum consectetur libero id faucibus. Ac turpis egestas maecenas pharetra convallis. Sagittis aliquam malesuada bibendum arcu vitae elementum curabitur vitae nunc. Nulla facilisi cras fermentum odio eu feugiat pretium nibh. Tortor at auctor urna nunc id cursus. Bibendum enim facilisis gravida neque convallis a cras semper auctor. Feugiat vivamus at augue eget arcu. Et netus et malesuada fames ac turpis egestas. Quisque id diam vel quam elementum. Amet est placerat in egestas erat. Egestas maecenas pharetra convallis posuere morbi leo. Sagittis aliquam malesuada bibendum arcu vitae. Ultricies lacus sed turpis tincidunt id aliquet risus. Ipsum dolor sit amet consectetur adipiscing elit. Cursus sit amet dictum sit amet justo donec.

17. Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Laoreet id donec ultrices tincidunt arcu. Sollicitudin aliquam ultrices sagittis orci a scelerisque. Sit amet aliquam id diam maecenas ultricies mi. Proin fermentum leo vel orci porta non. Ornare arcu dui vivamus arcu. Lorem ipsum dolor sit amet consectetur. Cras fermentum odio eu feugiat pretium nibh ipsum. Sapien nec sagittis aliquam malesuada bibendum arcu vitae elementum curabitur. Rhoncus est pellentesque elit ullamcorper dignissim cras tincidunt lobortis feugiat. Venenatis urna cursus eget nunc scelerisque viverra mauris in. Diam volutpat commodo sed egestas egestas fringilla phasellus faucibus. Sit amet volutpat consequat mauris nunc congue nisi vitae. Tincidunt ornare massa eget

egestas purus viverra accumsan in nisl. Semper quis lectus nulla at volutpat diam ut. Lobortis feugiat vivamus at augue eget arcu dictum varius duis. Vel facilisis volutpat est velit egestas dui id ornare arcu.

18. Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Tempor id eu nisl nunc mi ipsum. Gravida neque convallis a cras semper auctor neque vitae. In arcu cursus euismod quis viverra nibh cras pulvinar mattis. Pellentesque id nibh tortor id aliquet. Viverra adipiscing at in tellus integer. Volutpat lacus laoreet non curabitur gravida arcu. Arcu dui vivamus arcu felis bibendum ut tristique. Sollicitudin ac orci phasellus egestas tellus rutrum tellus pellentesque eu. Venenatis urna cursus eget nunc scelerisque viverra mauris in aliquam. Sociis natoque penatibus et magnis dis parturient. Morbi non arcu risus quis varius quam. Faucibus ornare suspendisse sed nisi lacus sed viverra tellus in. Sit amet commodo nulla facilisi nullam vehicula ipsum a arcu. Gravida in fermentum et sollicitudin. Aenean et tortor at risus.

Consequat ac felis donec et odio pellentesque diam. Nulla malesuada pellentesque elit eget gravida cum. Leo urna molestie at elementum eu facilisis sed. Nulla pharetra diam sit amet. Non arcu risus quis varius quam quisque id diam vel. Neque laoreet suspendisse interdum consectetur libero id faucibus nisl tincidunt. Platea dictumst vestibulum rhoncus est pellentesque elit ullamcorper. Velit laoreet id donec ultrices tincidunt arcu non sodales. Venenatis urna cursus eget nunc scelerisque viverra. Lectus magna fringilla urna porttitor rhoncus dolor. Proin libero nunc consequat interdum varius sit. Arcu felis bibendum ut tristique et egestas quis.

Refrences

[1].Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.
[2].Ipsum dolor sit amet consectetur adipiscing elit pellentesque. Orci eu lobortis elementum nibh. Faucibus a pellentesque sit amet porttitor.
[3].Egestas tellus rutrum tellus pellentesque eu tincidunt tortor. Sagittis orci a scelerisque purus semper eget. Vitae purus faucibus ornare suspendisse sed nisi lacus sed viverra.
[4].Augue interdum velit euismod in pellentesque massa placerat duis ultricies. Metus aliquam eleifend mi in nulla posuere sollicitudin aliquam ultrices.
[5].Velit laoreet id donec ultrices tincidunt arcu non sodales neque. Non curabitur gravida arcu ac tortor dignissim convallis aenean et.
[6].Euismod in pellentesque massa placerat. Morbi non arcu risus quis varius quam quisque.