DEVELOPMENT OF A ROBUST CAMERA-BASED TEXT RECOGNITION MODEL FOR THE VISUALLY IMPAIRED

BY

NWOKOMA, FRANCISCA ONYINYECHI

B.Sc, M.Sc [FUTO]

20164083208

A DISSERTATION SUBMITTED TO THE SCHOOL OF POSTGRADUATE STUDIES, FEDERAL UNIVERSITY OF TECHNOLOGY, OWERRI

DECEMBER, 2022

DEVELOPMENT OF A ROBUST CAMERA-BASED TEXT RECOGNITION MODEL FOR THE VISUALLY IMPAIRED

BY

NWOKOMA, FRANCISCA ONYINYECHI

B.Sc, M.Sc [FUTO]

20164083208

A DISSERTATION SUBMITTED TO THE SCHOOL OF POSTGRADUATE STUDIES, FEDERAL UNIVERSITY OF TECHNOLOGY, OWERRI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF PhD DEGREE IN COMPUTER SCIENCE

DECEMBER, 2022

CERTIFICATION

This is to certify that this work "**Development of a Robust Camera-Based Text Recognition Model for the Visually Impaired**" was carried out by **Nwokoma, Francisca Onyinyechi (20164083208)** in partial fulfillment of the requirement for the award of Doctorate (PhD) Degree in Computer Science, in the Department of Computer Science, Federal University of Technology, Owerri.

Dr. (Mrs.) J. N. Odii (Principal Supervisor)	14/11/2022 Date
Dr. I. I Ayogu (Co-Supervisor)	14/11/2022 Date
Dr. (Mrs.) C. G.Onukwugha (Co-Supervisor)	14/11/2022 Date
Dr. (Mrs.) J. N. Odii (Ag. HOD, Computer Science)	14/11/2022 Date
Prof. (Mrs) F.U. EZE (Dean SICT)	14/11/22 Date
Prof. B.O. Esonu. (Dean PGS)	Date
Prof. A. O. Adetunmbi (Eternal Examiner)	14/12/2022 Date

DEDICATION

This work is dedicated to Almighty God.

ACKNOWLEDGEMENT

First and foremost, I would like to thank my supervisors, , Dr. (Mrs.) J. N. Odii, Dr. I. I. Anyogu and Dr. (Mrs.) C. G. Onukwugha, for their selfless services, persistent guidance, encouragement and advice throughout the period of this study.

I would also like to express my profound gratitude to the Head of Computer Science Department Dr. (Mrs.) J. N. Odii for her dedication and commitment to the progress and Advancement of both undergraduates and postgraduate studies in Computer Science Department. To all the staff of Computer Science Department that contributed to the success of this work, am most grateful

My unreserved and endless gratitude goes to the coordinator of Post graduate studies in Computer Science Department, Dr. S. A. Okolie who in his wisdom and commitment has made postgraduate studies in Computer Science Department an interesting one. To all my Ph.D. course mates, especially my very own person, Mr. I. C. Nwandu, you are worth working with.

It will be an unpardonable offence if I fail to extend my heartfelt appreciation to my internal reviewer Dr. Anthony I. Otuonye. Sir, your corrections and suggestions were superb and they actually added to the success of this study.

To the Dean, School of Information and communication Technology, Prof. (Mrs) U. F. Eze, your effort and contribution to the advancement of this noble School cannot be overemphasized.

My thanks also goes to the Dean Postgraduate School Prof. C. C. Eze and his staff for facilitating post graduate studies and creating an avenue that allowed students to graduate as at when due. I would like to use this opportunity to express my profound and candid gratitude to the Vice Chancellor Prof. (Mrs.) Nnenna Otii and the entire management of this great and noble institution, Federal University of

Technology, Owerri for providing the funding which allowed me to undertake this research without financial stress.

Exceptional thanks also goes to the families of Prof. and Prof. (Mrs.) A. C. Onyeka & Prof. and Prof. (Mrs.) C. U. Njoku for their professorial Guides that birthed this huge success. Also to my most cherished brother in-law Mr. Ifeanyi Nwokoma, you are really wonderful and a rare gem.

These Recognitions will not be complete without my spiritual Father Ven. John Mere (the witness of Christ) whom the Lord through him has divinely ordered my steps and also my Spiritual Mother Rev. Elizabeth Sunday (mummy Lizzy), your prayers really went very far.

Special thanks goes to my parents, brothers and sisters and other family members especially Mr. and Mrs. Henry Uzochukwu, My Little Princess who has now grown to a responsible mother Mrs P. C. Ohanuka and her Husband Mr Kenneth Ohanuka, Mr. Alison Uzochukwu, my Kid Brother Anyanwu Daniel Maduabuchukwu and Oyibo Mustapha, I owe an endless gratitude. To my pretty damsels, Miss Glorious Nwokoma, Miss Delight Nwokoma, Miss Chibudom Nwokoma and their Aunty Miss Enyeribe Favour Adazion, thank you for tolerating and enduring all my scolding as a result of the stress and pressure associated to this work. Honestly, I wouldn't have asked for better blessing than you. You all are wonderful.

Special thanks go to my unique Husband, Mr. Innocent Uzoma Nwokoma for his Immeasurable Support, understanding, tolerance, care, and prayers in difficult times. You are truly Unique and amazing. In all, all Glory and Honour goes to the highest God who in his wealth of wisdom had made something out of nothing and has made all things beautiful in his own time.

TABLE OF CONTENTS

List of Figures

List of Tables

ABSTRACT

The quest to bridge the digital divide in this world of fast growing Information and Communication Technology should not only be restricted to some domains but should also be extended to all and sundry. Till date, Screen readers for the visually impaired still perform below expectation; their applications are also domain-dependent. Generally, research has shown that the Visually Impaired Persons (VIPs) tend to be greatly deprived of certain job opportunities due to their visual incapacitation and as such the unemployment rates among the visually impaired are increasingly alarming irrespective of their intellectual prowess. Therefore, to improve Text Recognition capabilities of OCR and incorporate the visually impaired community into employment setting, a Robust Camera Based Text Recognition model that will enable a blind person access documents and scene images for effective work collaboration is proposed. The system was designed to come up once the user machine is turned on. To bring this Concept to light, deep learning approach precisely CRAFT (Character-Region Awareness for Text Detection) Architecture which is suitable for detecting Curved images was deployed for text detection and CRNN (Convolutional Recurrent Neural Network) which combines the functionalities of CNN (Convolution Neural Network), RNN (Recurrent Neural Network) and CTC (Connectionist Temporal Classification) loss for an optimal Character Recognition was deployed. The Recognition Model was trained using Synth90k synthetic text dataset provided by the Visual Geometry Group (VGG) architecture which gives recognition accuracy of 98%. The system was implemented using Python Natural Language Processing Libraries. Finally, the recognized text is then communicated to the VIP in audio format.

CHAPTER ONE

INTRODUCTION

1.1 Background to the Study

The developments of vision assistance technologies have confirmed to elicit research attention. This is so because of the increasing positive social impart of vision assistant technologies on the blind population across the globe. Vision assistance technologies are specifically useful to the visually impaired persons (VIP) as well as the aged who cannot read without a carefully devised vision aid (Bhowmick, 2017). Hence, since accessing both print and soft document is an unavoidable part of live, reading print document and scene text by the visually impaired remains a viable research area (Lee, Lee, & Seo, 2021; Mishra, & Vedhapriyavadhana, 2022). Martin, (2016) in his study on "Technology Based aid for the visually impaired" stated that "the main issues for accessibility of a visually impaired or blind person are divided into six activities - mobility, communications and access to information, cognitive activities, daily living, education and employment, and recreational activities". Hence, the need for this research work which focuses on how to improve the quality of life of individuals with vision impairment specifically on the area of communication, access to information and employment activities with the aid of assistive technology. Braille, the foremost reading aid for the VIPs was invented in 1834 as a system of reading and writing by touch, and till date, it is still used as a means of accessing written materials. According to Valsan (2017), Braille by his invention rose above all odds to overcome the reading challenges imposed by vision loss.

The visually impaired and those with total blindness are fundamental part of every society. These communities of people have the potentials to contribute positively

to societal development. The VIPs, given the right support and opportunity will be of value to society, institution, government and economy just in a similar magnitude as any other member of the society. To address the challenges faced by the VIPs, especially with respect to gaining and maintaining employment, efficient, affordable and also adaptable assistive device, be it wearable or on wearable are required.

Despite the scale of research and development efforts towards the development of vision assistance technologies and tools to promote independent living of blind community, there are still significant researches, economic and practical challenges (Nadhir *et al.*, 2021;). Available vision assistive devices are either too sophisticated or outside what the average, poor VIP can afford, thus leaving the indigent VIPs to fate.

Screen readers still perform poorly on text images sandwiched in a Complex, noisy background, images with poor lighting features and texts with regular font sizes and styles; their applications are also domain-dependent (Lazar *et al.*, 2007; Kuber & Amanda, 2012). This situation is attributed to the challenges of unrestricted text recognition which include complexity of environments flexible mage acquisition styles and variation of text contents (Ye & Doermann, 2015 ; Kantipudi *et al*., 2021). Text detection and recognition in documents with multiple parts, complex and heterogeneous background, mixed fonts or other irregular characteristics is much more challenging than in regular, standard font, plain printed text documents (Gupta & Banga, 2012; Shikha *et al*., 2020). The difficulty of this task is further aggravated when "the document image is poorly captured either due to operator error, incompetence, low device quality or poor environmental conditions such as lighting when a camera is used to capture the image" (Mhaske & Sadavarte, 2016; Trémeau *et al*., 2011).

Capturing and converting scene images to text can be best handled by hand held or wearable camera-based OCR rather than the traditional flatbed desktop OCR scanner. This is obviously so because desktop scanners are not designed for such applications. Therefore, considering document and scene text image analysis, detecting and recognizing text have several practical applications other than vision assistance. Scene text recognition has application in content based image retrieval automatic navigation systems in vehicles and digitization of text books (Mahmood & Edirisinghe, 2017; Mahmood, 2020; Kantipudi *et al*., 2021).

Globally according to WHO 2022, there are about 2.2 billion vision impaired people amongst them are those with good educational background and relevant qualifications for employment. But majority have remained unemployed. This is in spite of the fact that if employed, this population can contribute novel ideas that could increase the values of organizations. According to Fawaz *et al.*, (2013), the blind communities considers themselves as a burden to the society and they do not engage in basic routine activities which results in isolating the blind people from the society". The fact remains that vision impairment is never a curse neither is disability the same thing as disqualification. One might be physically blind but not cognitively blind. It is recognized that "inclusion in the workplace is critical to full participation in society and to financial independence. The VIPs have historically been under-represented in the labour market.

This research therefore proposes to explore the first order edge detection (canny edge detection) techniques and the deep learning approach for the improvement of character recognition accuracy in regular, multipart document, such as an official memorandum and in scene images using a hand-held camera and the recognized text is communicated to the visually impaired via audio format. This marks the uniqueness of the proposed research and thus sets up the trajectory for realizing a

system that can assist the visually challenged in reading texts from document and scene images using a hand held or wearable camera.

1.2 Problem Statement

Text detection in documents with multipart or complex background can be more challenging than extracting text from a normal document with regular fonts and size. Detecting text in Multi-parts document imposes lots of computational difficulties as encountered in a letter headed document or an office memo where text can assume different fonts, resolutions, irregular arrangement and some may also overlap (Fei *et al*., 2018) . Secondly, since text localization is a Camera-Based OCR, camera setting may also results to blur images or perspective distortion thereby leading to text degradation. The amount of light available within the text/document environment also affect the quality of the captured image. When the lightening is insufficient, it creates reflection on the document surface. Objects around the document may also cast shadow on the document. Hence, the computational complexity is increased and blurred text is produced. Therefore, for a camera based OCR, the photometric effect has to be reduced and the sigma's function has to be very small in order to increase the brightness of the captured text. Defining a region of interest as proposed by (Karande & Jogdand, 2018) also improve performance and text readability. According to Valsan (2017), Braille is the most famous communication system for the blind people but it is cumbersome to use and difficult to learn.

1.3 Objectives

The Primary objective of this research is to develop a robust text recognition model for the improvement of camera-based reading for the visually impaired. The specific objectives are to:

i. develop text detection model for camera-based OCR in documents and scenes using image enhancement, text localization and dimensionality reduction techniques.
ii. develop text recognition model based on objective (i) above.
iii. evaluate the models developed in (i) and (ii) using appropriate datasets
iv. develop a prototype camera-based reading system by integrating the models developed in (i) and (ii) into an off-the-shelf, open-source speech synthesis software
v. evaluate the prototype developed in (iv) using human subjects.

1.4. Justification of Study

Generally, the VIPs tend to be highly neglected in terms of employment. They are severely confronted with challenges like being treated as outcast, people without value and the minority group. This misconception and negative references can cause individuals in this category to "lose self-esteem, personal satisfaction and hinder them from obtaining an appropriate job for self-governing life style" (Fawaz *et al*., 2013). Though, so many assistive technologies have been designed to enable the visually impaired persons work comfortably in an office, these devices sometimes are not available, affordable and accessible to the targeted end users. Hence, the need for an improved, easy to use and affordable system.

As noted by Fawaz *et al*., (2013), "blind community considered themselves as a burden to the society and they do not engage in basic routine activities which results in isolating the blind people from the society"; this system will incorporate them into the society and eliminate the barrier of isolation and negative perceptions as a result of misconception. Therefore, to improve Text Recognition capabilities of OCR and incorporate the visually impaired community into employment setting, a Robust Camera Based Text Recognition model that will enable a blind person access documents and scene images for effective work collaboration is proposed. The system was designed to come up once the user machine is turned on. To bring this Concept to light, deep learning approach precisely CRAFT (Character-Region Awareness for Text Detection) Architecture which is suitable for detecting Curved images was deployed for text detection and CRNN (Convolutional Recurrent Neural Network) which combines the functionalities of CNN (Convolution Neural Network), RNN (Recurrent Neural Network) and CTC (Connectionist Temporal Classification) loss for an optimal Character Recognition was deployed. The system if successfully implemented and adopted will enhance the reading ability of the VIP and as well give them the opportunity of being gainfully employed and working comfortably in an office. There by resuscitating their self-esteem, personal satisfaction, and self-dependent. This research also, "has the potential to identify novel ways that vision-restricted individuals obtain and maintain employment" thereby justifying the claim that employment is an occupational right as noted by (Mohler, 2012)**.**

On the contrary, failure to implement this system will continue to increase the low esteem and total dependency of the VIPs upon relatives. Again, it will as well deprive them the opportunity to participate fully and contribute their own quota toward the development and advancement of their country's economy.

1.5. Scope of the Study

Evidently, many positive technological advances had been made by both government agencies and renowned researchers on the area of improving independent living of blind community, yet their challenges has not been totally taken care of. Text detection and recognition in documents with multiple parts, complex and heterogeneous background, poorly captured images due to (operator error, incompetence, low device quality or poor environmental conditions such as lighting when a camera is used to capture the image), etc., mixed fonts or other irregular characteristics is much more challenging than in regular, standard font, plain printed text documents which the traditional desk OCR is very comfortable with (Gupta & Banga, 2012; Mhaske & Sadavarte, 2016; Trémeau *et al*., 2011; Mahmood, 2020; Shikha *et al*., 2020).

Therefore, this research is set out to develop a camera-based OCR system for documents and scene Text reading. This is achieved by developing both text detection and recognition model for camera-based OCR and evaluating the models using appropriate datasets. After which, a prototype camera-based reading system which convert the recognized text to speech and communicate to the VIP via audio format is developed. The developed system should be enabled to run automatically as soon as the user system is turned on.

CHAPTER TWO

LITERATURE REVIEW

2.2 Artificial Intelligence

Generally, Artificial Intelligence (AI) is defined as a subfield of Computer Science that attempts to develop computer systems that can mimic or simulate human thought processes and actions. AI has also been defined as the study of ideas that enable a computer to be intelligent (Ugwu *et al.*, 2012). Habeeb (2017) defined AI as an "area of computer science that emphasizes the creation of intelligent machines that work and react like humans". Even with the fast advancement in the machine industry, intelligence has remained the major difference that exists between humans and machines. Human beings are able to gather information from their surroundings using the various senses: vision, hearing, smell taste and tactility. This information is analyzed, processed and suitable decisions are taken with the help of the brain. On the contrary, machines are not ordinarily intelligent.

According to Mohammed, & Bashie, (2016), "the era of intelligent machines started in the mid-twentieth century when Alan Turing thought whether it is possible for machines to think. He demonstrated this using a machine game. In 1950, Alan Turing "published an essay in which he poses the question whether machines can think. Because of the ambiguity of the terms *machine* and *think* he suggests an empirical test, in which a computer's use of language would be the basis for determining if a machine can think. In this test, or game, nowadays known as the Turing test, there are three participants. Two of them are human and one is a computer. One of the humans plays the role of an interrogator with the objective of determining which of the other two participants, is the computer. He

or she has to do so by asking a series of questions via a teletype. The task of the machine is to try to fool the interrogator into believing it is a person. The second participant also wants to convince the interrogator that he or she is human. Because of the fact that effective use of language is intertwined with our cognitive abilities, Turing believed that using language as humans do is sufficient as an operational test for intelligence" (Woudenberg, 2014).

Since then, the artificial intelligence (AI) branch of computer science has developed rapidly". Among the most popular application of AI include: speech recognition, learning, planning, problem solving etc". Figure 2.1 proposed by Chethan, (2018) shows the various ways of achieving AI (branches of AI).

But for the purpose of this study, Machine Learning, Computer Vision (image/ face recognition) and Speech (text to speech) conversion shall be concentrated on.

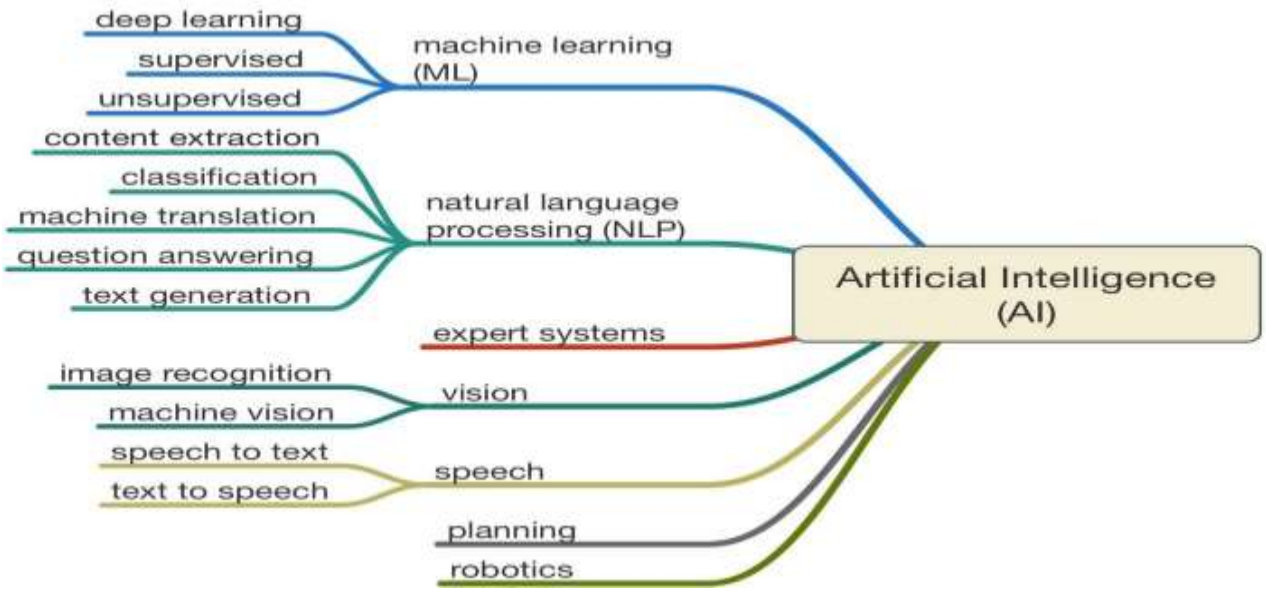


Figure 2.1: Branches of AI

2.2.1 Machine Learning (ML): Machine learning is a branch of artificial intelligence that aims at enabling machines to perform their jobs skillfully by using intelligent software (Mohammed *et al.*, 2016). It is used to "teach machines how to

handle data more efficiently. In most cases, after viewing data, it may be difficult to interpret the pattern or extract information from the data. In that case, machine learning is applied" (Dey, 2016). In this approach, "the target is defined and the steps to achieving that target are learned by the machine itself by training (gaining experience)". Hence, the purpose of machine learning is to learn from the data. Humans most of the times perform difficult task without efforts, but find it very difficult to explain how the task was carried out. For instance, a child can easily recognize the voice of a close relative without much difficulty. Machine learning algorithms are helpful in bridging this gap of understanding. The idea is very simple. The target is not to understand the underlying processes that help humans learn. Instead, the computer programs that will make machines learn and enable them to perform tasks, such as prediction are written" (Mohammed *et al.*, 2016). The term machine learning "means to enable machines to learn without programming them explicitly. There are four general machine learning methods: (1) supervised, (2) unsupervised, (3) semi-supervised, and (4) reinforcement learning methods".

For the purpose of this study, the supervised machine learning approach is adopted. The supervised machine learning algorithms "are those algorithms which needs external assistance. The input dataset is divided into train and test dataset. The train dataset has output variable which needs to be predicted or classified" (Dey, 2016). The goal of this learning approach is often to get the computer to learn a classification system that has been created.

2.2.2 Computer Vision: This is an area of AI study that enables the computer to efficiently derive meaningful information from digital images, videos and similar data. Computer vision according to Martin, (2016), takes on an algorithmic approach to analyzing and extracting information from images. It captures and

analyses visual information using a camera, analog-to-digital conversion, and digital signal processing. For example, according to Microsoft (2018), “Seeing AI is a free app from Microsoft that will narrates the world around you. It is designed for the low vision community, and it harnesses the power of AI to describe people, text and objects using computer vision”.

2.2.3 Speech Processing: Speech Recognition as documented by Trivedi *et al.*, (2018) is the ability of machine or program to identify words and phrases in spoken language and convert them into machine-readable format. These Systems can be classified based on the following parameters:

a. Speaker: All speakers have a different kind of voice. Hence, the models are designed either for a specific speaker or an independent speaker.
b. Vocal Sound: another factor that plays a big role in speech recognition is the way the speaker speaks. Some models can either recognize single utterances or separate utterance with a pause in between.
c. Vocabulary: in determining the complexity, performance, and precision of the system, the size of the vocabulary plays a vital role (Trivedi *et al.*, 2018)”.

In Text-To-Speech conversion, input text is first analyzed, then processed and understood, and then the text is converted to digital audio and then speech. Figure 2.2 shows the block diagram of Text-To-Speech system and the basic steps involved in the text to speech conversion as mentioned in Trivedi *et al.*, (2018)

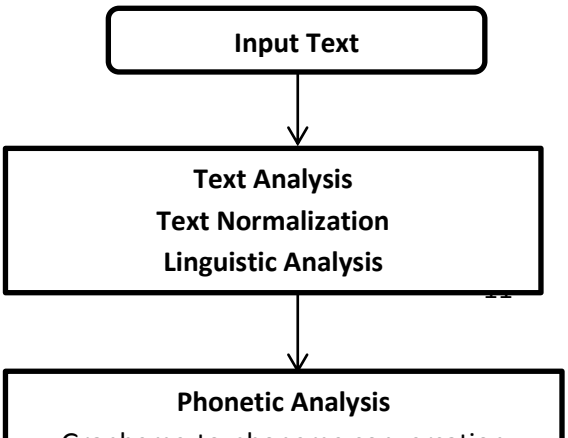


Figure 2.2: Block diagram of text-to-speech system.

2.3 Technology Aid for the Visually Impaired Persons

A research field that is attaining a persistent Popularity is Assistive technology for the visually impaired and blind. The field has a very relevant social impact on the blind populations and ever-increasing aging population (Bhowmick, 2017). An assistive technology device is therefore any item, piece of equipment, or product system, whether acquired commercially off-the-shelf, modified, or customized, that is used to increase, maintain, or improve the functional capabilities of individuals with disabilities. Assistive also called adaptive technologies are electronic solutions that enable people with disabilities to live independently.

Blind persons can hear computer-screen text, and people with visual impairments can enlarge text, enabling independent reading. People who are unable to manipulate a mouse can enter data. Assistive technology services are any services

that directly assist an individual with a disability in the selection, acquisition, or use of an assistive technology device.

These assistive technologies range from early, low, mid and high tech assistive technologies. Eyeglasses remain the most commonly low tech device for vision correction. To increase the size of text, images, or objects for people with visual impairments, Magnifying glasses are used. To access printed materials, Braille, large print materials, and audio versions of materials such as books on tape are used. For people with visual impairments, navigating environments can be difficult. But there are low tech devices that can support in this. These include the Hoover Cane, the white cane used in navigating environments to feel upcoming obstacles in order to avoid them, and Seeing Eye dogs which are used in a similar manner.

Another category of assistive technology is the Mid Tech. mid tech devices are usually household devices equipped with an audio output. Examples of devices in this category include talking calculators, talking watches, talking clocks, and talking thermometers that announce times or numbers.

Finally, the high rate of technological advancement has also led to evolution of high Tech assistive technology. This technology primarily consists of computer or printed material access options. Computer access options include “Braille keyboards, screen magnification software, screen readers, and text-to-audio programs. These devices provide people with visual impairments access to a computer by reading the contents of a computer screen to aid in navigation, magnifying the size of objects on the screen to an optimal viewing size, or reading letters that are typed to ensure accuracy. Another high tech computer access device is the Power Braille, a device that connects to a keyboard and prints what is being

typed both on the computer and a refreshable Braille display. Assistive technology therefore is aimed at making the world more accessible to people with vision restriction as to increase their standard of living.

2.3.1 Wearable Devices

Wearable technology as noted by Babu *et al*., (2015) is a group of electronic devices that can be worn as accessories, embedded in clothing, implanted in the user's body, or even tattooed on the skin. These devices are hands-free gadgets with practical uses, powered by microprocessors and enhanced with the ability to send and receive data via the Internet. Wearable assistive devices represent potential aids for people with physical and sensory disabilities that might lead to improvements in the quality of life. The body areas that are involved in wearable assistive devices includes: fingers, hands, wrist, abdomen, chest, feet, tongue, ears, head etc. examples of wearable devices are Maptic, Horus, OrCam, etc

2.3.2 Assistive Technology for the VIPs

Before the invention of non-visual aid technologies, the two main tools for obstacle avoidance remain the long cane and guide dogs. “The long canes can only feedback information about the area up to one stride ahead of the user. This can sometimes be increased by tapping the cane and calculating a frame of reference by analyzing the echos of the sound. The long cane is limited in its accuracy and resolution and cannot be used effectively for navigation purposes - for calculating the best route to a location” (Martin, 2016). In this sections, different assistive technology for the VIP are discussed as follows

1. **Laser Cane**: the two main types of Laser can are the Smart Cane and Ultra Cane. To detect obstacles in the way of the users, these devices use ultrasonic

echoes which are embedded in the handle as documented by Wahab et al., (2011) in figure 2.3 to offer spatial awareness to the VIP.

 .

Figure 2.3: Smart cane.

2. **BlindSquare and Seeing Assistance**: BlindSquare is an assistive technology that uses the latest features available in smartphones to help the blind and VIP in their daily lives. With "iOS-device's and GPS capabilities, BlindSquare determines your location and looks up information about your surrounding on Foursquare and Open Street Map".

3. **Sentiri:** is a head band that detects obstructions in all directions, not just straight ahead. It "consist of a 360-degree headband that uses IR sensors to detect obstacles and inform the user where the closest obstacles are using haptic feedback on the head". The diagram in figure 2.4. by Atwell, (2015) shows how to place Sentiri for proper functioning and better performance.

Figure 2.4: Sentiri.

4. **Wayfindr:** help VIP to navigate through busy environments like train stations, via audio output. The installed beacons relay the information via Bluetooth toWayfindr users in the vicinity.

5. **OrCam:** allow the user access objects or information in the scene around them Using vision technology. It has a very high quality Camera that is fixed to a pair of glasses, large memory capacity OCR technology for scanning and converting print document to text and then to speech. According to the OrCame CEO Aviram, (2018), OrCam is "the most advanced wearable device in the world that can assist visually impaired and blind people. It can read text, recognize faces, money notes, credit cards identify millions of products. OrCam MyEye 2 is easy-to-use and extremely intuitive, with users as young as 7 and as old as 100. It is available in 29 countries, in nearly 20 languages, including English, Spanish, German, Italian, Dutch, French, Norwegian, and Polish. Figure 2.5 as indicated by Aviram, (2018) shows where to position the Orcam on a pair of glasses and how it works. Orcam has virtually all the features needed for a blind person to work effectively and live independently, but it major challenge is the cost of acquiring it which is $3,500, approximately 3million in Naira equivalent.

Figure 2.5: OrCam Attached to a Pair of Glasses.

6. **Screen Enlargement Programs**: use software to enlarge print on a computer screen from 2X to 16X. Examples of sreen magnifiers include: MAGic, Windows Magnifier, Zoom on Macs, and ZoomText (AFB, 2003).

7. **Optical Character Recognition (OCR)**: This is one of the leading branches of Pattern Recognition system. As noted by Shikha et al., (2020), "OCR is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text. OCR technologes enables character recognition with high accuracy. It involves scanning a printed document into a computer and converting the picture image into text characters and words, which screen readers and braille embossers can recognize. There, the adaptive program converts the image into computer text and reads the information out loud".

8. **Horus**: this comprises a band that wraps around the back of the head; with earpieces; and two side-by-side cameras on one end. The cameras observes what's in front of the wearer, and can dictate what it sees through the earpiece, which uses bone conduction technology to bypass the ear canal and stimulate the tiny ear bones directly. figure 2.6 documented by Martin, (2016) shows the different part of horus. Siri by apple, Google Assistant by Google, and Alexa by Amazon perform almost the same function with Horus.

Figure 2.6: Horus Device.

9. **Tactile**: Tactile Developed by TeamTactile make printed text more accessible to the blind. Some of the printed material has braille translation and as a result , simple tasks, like reading of mail becomes inaccessible. Therefore, in an attempt to make those braille materials accessible, Tactile Team as noted by Teo, (2018) developed this device in figure 2.7 that glides over printed pages, scans the text, and translates it into braille, line by line.

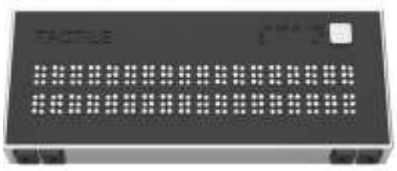


Figure 2.7: Tactile Device.

10. **Maptic**: Maptic was designed by Emilios Farrington-Arnas. It is like "a fitness tracker, Consisting of a visual sensor, which is worn around the neck, and vibrating feedback units worn around the wrists; the Maptic system discretely helps visually impaired users navigate the world around them". Fiure 2.8 shows diagram of Maptic device as noted by Teo, (2018). Through haptic feedback, users are guided through a physical environment.

Figure 2.8: Maptic Device.

11. **A Speech Access System**: According to AFB, "speech access systems are software programs that enable individuals who are blind or visually impaired to access text and graphics on the computer screen by magnification, speech output, and/or braille output to a refreshable braille display. These assistive technologies are examples of software/devices that can help those who are blind or visually impaired read printed material or surf the web". There are different types of screen readers:

i. COBRA (Baum): works with Microsoft Windows operating systems. It allows for magnification up to 32x and provides braille output. There are three versions available: "COBRA Braille, COBRA Zoom, and COBRA Pro. COBRA Braille offers braille and speech output, COBRA Zoom offers screen magnification with speech output, and COBRA Pro offers braille output, screen magnification, and speech output. Both COBRA Zoom and COBRA Pro models provide full- or split-screen viewing, as well as optical aids such as cursor highlighting, focus border, and mouse pointer. Cobra Braille also handles multiple braille displays connected by USB or through Bluetooth". Other Screen readers include Job Access to Speech (JAW, for windows), Narrator (for windows), Non-Visual Desktop Access (NVDA), Kurzweil Education.

Finally, the skills of Touch-typing eliminate the need to look at the keyboard while typing. It is one of the important and useful skills blind and virtually impaired person can learn. Touch-typing opened up new career opportunities as blind people

can now work comfortably as a typist. Thanks to today's technology, being blind or virtually impaired no longer place the many restriction on independence. With these technologies in place, individuals who are blind or virtually impaired can live on their own, receives quality education and pursue their chosen career path.

2.3. Vision Impairment

An impairment according to (Martin 2016), can be of two main approaches: the medical and the social models. The medical model is defined by the World Health Organization (WHO) as: any loss or abnormality of psychological, physical or anatomical structure or function. The social model is defined by the Disabled People International (DPI) as: The functional limitation caused by physical, sensory or mental impairments.

2.3.1. The Magnitude of Vision Impairment

According to a fact sheet put forward by the WHO on 13th October 2022, global estimate of VIPs is at least 2.2 billion people. At least 1 billion people have a vision impairment that could have been prevented or has yet to be addressed. This 1 billion people includes those with moderate or severe distance vision impairment or blindness due to unaddressed refractive error, as well as near vision impairment caused by unaddressed presbyopia.

Globally, the leading causes of vision impairment are uncorrected refractive errors (123.7 million), cataract (65.2 million), glaucoma (6.9 million), corneal opacities (4.2 million), diabetic retinopathy (3 million), and trachoma (2 million), as well as near vision impairment caused by unaddressed presbyopia (826 million). In terms of regional differences, the prevalence of distance vision impairment in low- and middle-income regions is estimated to be four times higher than in high-income

regions. Majority of people with vision impairment are over the age of 50 years". Some of the other conditions that fulfill the criteria for Visual Impairment are:

1. Photophobia: This is inability to clearly see in light,
2. Diplopia: where an individual has double vision,
3. Visual Distortion: where images of objects feel distorted,
4. Visual Perceptual Difficulties: where an individual finds it difficult to gauge the depth or length of an object.

Blindness is a form of visual impairment in which the visual acuity of an individual is extremely poor along with the visual field where the individual is not able to see any object.

2.3.2. Visual Impairment as an Imperative Global Health Issue

Total blindness and visual impairment are overbearing global health issues because of the following reasons:

1. They lead to increased indisposition and mortality,
2. Decreased quality of life.
3. They lead to a substantial economic productivity loss.

The severity of the above 3 listed point under visual impairment as a global health concern have really drawn so many researchers attention to inventing new technologies and ideas that will help boost the self-esteem of these class of individuals and also make them feel digitally, socially, emotionally and generally included. Still in the quest of this general inclusion of the blind community,

Sotomayor *et al.*, (2019) proposed and designed a system to help visually impaired people to navigate safely and quickly between obstacles and other dangers facing their movement. Considering the intensity of mobility issues confronting the visually impaired, Sahoo & Chang, (2019) came up with a system comprising of walking stick and Android-based applications that will aid them in detecting and avoiding obstacles and water puddles in their way. Still on mobility issue, Borenstein & Ulrich, (1997) design a device called GuideCane, to enable vision impaired navigate both In indoors and outdoors. Mahmud, Saha, & Islam, (2013) build a Smart walking stick to enhance the free movement of the VIP. Their system is very similar to what Borenstein did with the same limitation of detecting obstacles only in front of it. Noman & Rashid, (2017) sees blindness as a more severe problem among other disabilities of human. having noted this, they went into research and discovered that most of the devices invented by several researchers to aid independent navigation for the blinds are for a particular task and also not cost friendly. So, to come up with a universal and a more cost friendly device, Noman *et al.,* (2017) designed and implemented a Blind Assistive Robot that eliminates the requirement of human assistance of the blind while traveling outside. Still on the mobility enhancement for the blind community, Martinez-sala *et al*., (2015) having perceived indoor navigation a great challenge to the visually impaired persons, developed an indoor navigation system called SUGAR that uses an Ultra-Wideband (UWB) technology which is proven to be more effective in terms of indoor positioning. Yet on the issue of improving the independent living of the blind and visually impaired counterparts, Vaz *et al.*, (2018) noted that "blind and visually impaired visitors experience a lot of constraints when visiting museums exhibitions giving the ocular centricity of these institutions, the lack of cognitive, physical and sensorial access to exhibits or replicas, increased by the

inaccessibility to use the digital media technologies designed to provide different experiences, among other constraints. On this note, they design and implemented an exhibitor to communicate original museum samples to blind and visually impaired patrons, without the need to be replicated, that interactively "tell stories about their lives whenever picked up

2.3.3. Employment after Vision Loss and Basic HCI for the Blind.

Considering employment after vision loss, Crudden, (2002) Ventured into research by deploying a collective case study approach to examine Ten participants on how job modifications, job restructuring, training/retraining, organized labor, transportation, motivation, and their employers and coworkers impacted their job retention after vision loss. Among other challenges confronting the VIP, print access and technology were a source of stress for most participants. However, the sample pool is not random and is not representative of the general population of persons who are visually impaired or who retained employment after vision loss.

The research work on Process Of Obtaining And Retaining Employment Among The Vision-restricted by Mohler, (2012) used a constructivist Grounded Theory methodology to gain an in-depth understanding of what people with vision-restrictions, who perceived that they were successfully employed, considered to be fundamental in the search for employment. However, his interview was not extended to the employers to know why some of the identified barriers still persist and what exactly is preventing the removal of these obstacles. He focused only on the identification of possible barriers confronting the virtually impaired community, and the possible solutions to the identified challenges were not implemented.

Raymond & Alcides, (2010), having worked with blind people for almost two years, he observed some of the obstacles they need to overcome to complete their school. On these bases, he went into a research on basic human computer interface for the

Blind. His intention was to "design a blind person's user interface based on their behavioral characteristics and provide them an independent and enjoyable environment". Kulyukin & Kutiyanawala, (2010) and López-de-ipiña *et al*., (2011) carried out a research to identify the several design requirements for assistive shopping systems and analyze existing approaches to see how well they meet them.

Duquette & Baril, (2013); Hussin, (2013) carried a review study on the factors influencing work participation for people with visual impairment. In their study, they were able to identify non personal modifiable factors like severity of vision loss, age at the time of vision loss, multiple impairment, sex, etc. and personal modifiable factors such as behavior, education, mobility, employment and responsibility etc.

Donnell, (2014) carried out research to investigate whether or not they were employed and the types of barriers they faced while they looked for employment. His findings revealed continued negative trends in employment rates amongst blind people, attitudinal and technological barriers and lack of protection or advocate for the blind. On the same note, Hanley, (2014) embarked on a research to Conceptualize disability in the workplace. His major focus was on one element of the employment system, which is how managers and co-workers conceptualize disability and disabled workers, exploring the likely impact this could have on sustainability of employment for disabled people".

Doughty, (2016) carried out a research to ascertain the level that assistive technologies can aid people with one form of disability or another. In his search, he was able to identify as many assistive technologies as possible, like Field cameras, peep-hole viewers and linked internal cameras for improved security. Nadin, (2017) carried out a research on a phenomenological exploration of visually impaired

individuals' perspectives on obtaining and maintaining employment. This was further investigated by (Pruettikomon & Louhapensang, 2018; Crudden, 2018).

2.4. Camera-Based-OCR

Camera-based OCR technology represents a set of related methods developed to recognize texts in digital images or Video. Camera-based OCR has increased in popularity partly due the emergence of portable digital image processing device such as mobile phones, cheap-hand held camera. The ease with these modern devices can be used to process scene images was not previously obtainable with the tra`ditional flat-bed desktop scanners. Hand-held digital cameras such as high-end cell-phones, Personal Digital Assistants (PDA), smart phones, iPhones, iPods, etc possess a very high processing speed and internal memory which encourages real time processing" (Mollah *et al.,* 2011). Though, "characters captured by camera are different from those in scanned documents and their recognition is very difficult" (Uchida, 2020). Text detection and recognition in documents with multiple parts, complex and heterogeneous background, mixed fonts or other irregular characteristics is much more challenging than in regular, standard font, plain printed text documents (Gupta & Banga, 2012). The difficulty of this task is further aggravated when the document image is poorly captured either due to operator error, incompetence, low device quality or poor environmental conditions such as lighting when a camera is used to capture the image (Mhaske & Sadavarte, 2016; Trémeau *et al.,* 2011), because of the non-contact nature of digital cameras attached to handheld or wearable devices, the output of acquired images mostly suffer from skew and perspective distortion.

2.4.1. Tasks in Camera-Based OCR

Text recognition in scene images, multipart document, documents with complex background and video comprises three major tasks: Image acquisition/data collection, text localization and text Recognition. These tasks are as shown by Uchida, (2020) in figure 2.9.

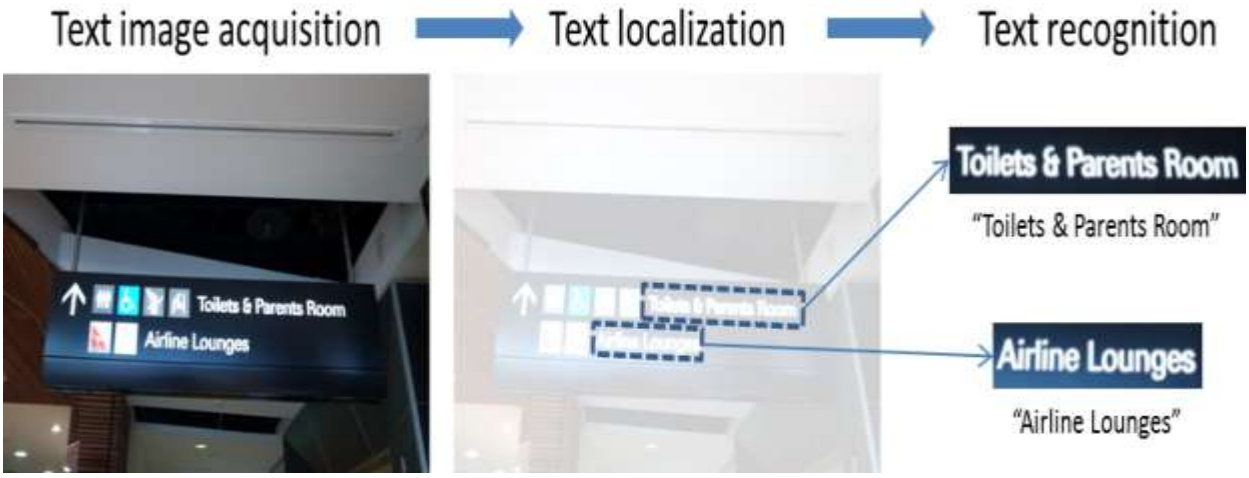


Figure 2.9: Camera-Based OCR Tasks.

a. Image Acquisition/Data Collection:

The process of detecting and recognizing text in scene images begins with the collection of appropriate data. Primarily, image data is acquired using an imaging device of appropriate quality. This process is of great importance to the entire process. As noted by (Uchida, 2020), Many difficulties of camera-based OCR such as low resolution, motion blurring, perspective distortion, non-uniform lighting, specular by flash, and etc are caused by this acquisition step. To ensure these issues are avoided, the following steps should be taken:

1. To acquire high-resolution images at some distance, we need to perform zooming to the area of interest (Doermann & Liang, 2003).
2. To avoid low resolution, do not capture all the text in one frame.
3. The mosaic technique is needed to put pieces of text images together as a large high-resolution image.
4. Auto zooming should be used in the background around an object that has low variance compared to the main object. In this case, the variance in the observation window could be used as an indicator of best zoom (Mirmehdi & Palmer, 1997).
5. The best focus is achieved when edges are strongest in the image.
6. The sum of the differences between neighbouring pixels, the sum of gradient magnitude, or the sum of a Laplacian filter's output could be used as the overall edge strength measure.
7. To avoid perspective distortion, the camera should directly face the document (Zandifar, Duraiswami, Chahine &, 2002).

b. Preprocessing:

The main objective of preprocessing is to create a distinction between the various aspects of input image. Image preprocessing is primarily conducted to establish a focus on the area of interest in a given mage. This enables the efficient processing of the features of interest leaving the algorithm with more specific data devoid of noise and other forms of interference.

A variety of techniques have been developed by research for image preprocessing that are of relevance and significance to the problem of scene text recognition. A few of this include binarization, threshholding, Skew correction etc.

i. Binarization: converts a full coloured image image into a gray scale image. As a basic rule, binarization can be done by fixing a threshold, normally threshold=127, as it is exactly half of the pixel range 0–255. In this scheme, pixel value greater than the threshold value is considered as a white pixel, else it is considered a black pixel (Sunsmith, 2019). This techniques is noted to fail when lightening are not Uniform in the image as shown in Figure 2.10.

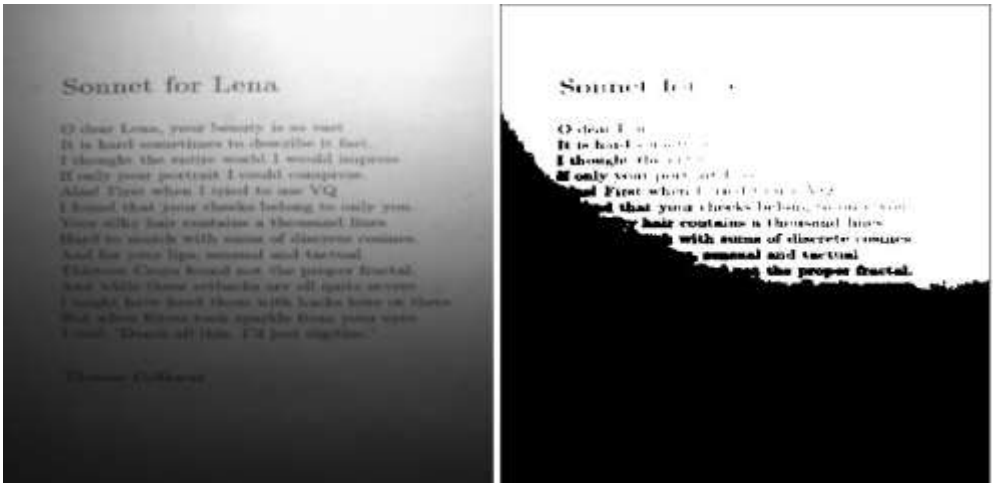


Figure 2.10: Effects of Lightning on Binarization

This means that the fundamental issues in binarization revolve around determining the threshold. To address these challenges research has establish a number of techniques. These are briefly outlined next.

ii. Local Maxima and Minima Method: This method uses contrast which depends upon the local minimum and maximum. A normalization factor is introduced which will compensate the effect of variation in image background. Image contrast is calculated as shown in equation 1. (Chauhan *et al.*, 2016)

$$C(i,j) = \frac{l\max - l\min}{l\max - l\min + \mathrm{E}} \quad (1)$$

Where: I_{max}= maximum pixel value in the image, I_{min}= minimum pixel value in the image, E = constant value.

iii. Otsu's Binarization: Otsu's method is a global thresholding method that converts a gray scale image into a bi-level image. This technique divides the pixels into two classes one is foreground and the other is background. It chooses an optimal threshold that separates the image into two different classes. Otsu method gives its best performance for images that have clear bi-modal pattern. It dose not Perform well on degraded documents that lack clear-cut pattern (Smith *et al.,* 1979; Chauhan *et al.,* 2016). The basic steps in Otsu's algorithm are outlined in algorithm below:

Step 1: Separate the pixels into two clusters according to the threshold.
Step 2: Find the mean of each cluster.
Step 3: Square the difference between the means.
Step 4: Multiply by the number of pixels in one cluster times the number in the other. (Poornima et al., 2011)

Algorithm: Otsu's Binarization algorithm

iv. Adaptive Thresholding: Adaptive Thresholding merges image contrast and image gradients. Contrast map is constructed and binarized. Canny edge map is also constructed and combined with the resultant of the first step. This detects the actual text stroke edges. The text is extracted using local thresholding. It is estimated from the mean and standard deviation of the detected text stroke pixels within a window. The limitation of this particular

method is that the canny edge detector extracts the fake edges also (Su *et al*., 2012; Chauhan *et al.*, 2016)

c. Text Localization:

This is the second major step in the process of scene text recognition. It involves text detection in a scene image, document with multipart and video frame. It ascertains the existence of text in the captured image(s). Research has shown that there are a number of challenges associated with localizing text in scene images especially those with complex background, poor lighting among others. Some of these challenges are briefly outlined next.

a. Scattered Text and words in the scene image with no previous information on text orientation, location, size, and the number of texts. Because there is no formatting rule in scene text, localization cannot be done by text segmentation methods for document images.

b. Complex Background: Text printed on a background that has a very solid colour, separation is rather easy like ordinary scanned documents. In scene Images, characters and their background often have a very low contrasts and as such, separation is a little more difficult even when a solid colored background is used. Captions are mostly without solid background because they are placed on top video frames directly.

c. Pseudo Text Patterns: There are many patterns which seem like characters as a result of confusing and ambiguous shapes in the scene. “Y” shaped edges are found on the corner of a room. Around leaves of trees, we also find dense and fine (i.e., high-frequency) edge structures which look like dense text lines. Equally, “it is also very obvious that some decorated characters seem like a

branch of a tree. Normally, scene images are often curved, rotated or even slanted.

There are so Many existing approaches and methods proposed by different researchers for retrieving text from scene images and document with multipart and complex background. Figure 2.11 is the diagram of the steps in Text Information Extraction as documented by (Yadav & Kumar, 2016)

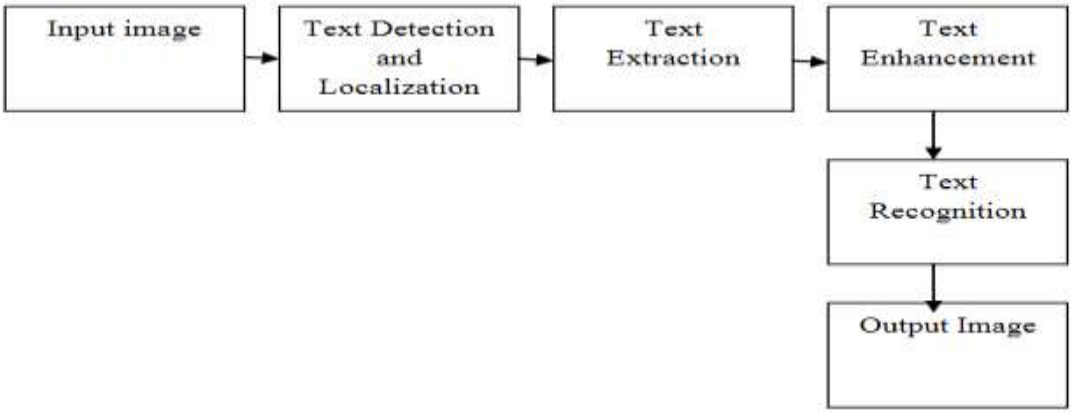


Figure 2.11: Text Information Extraction Model.

2.4.2 Text Localization Techniques

Text Localization Techniques are broadly classified into Region based, texture-based and hybrid approaches. The region based methods are further classified into connected component based and edge based. These approaches were documented by Chavre, (2015); Yadav & Kumar, (2016) as shown in figure 2.12

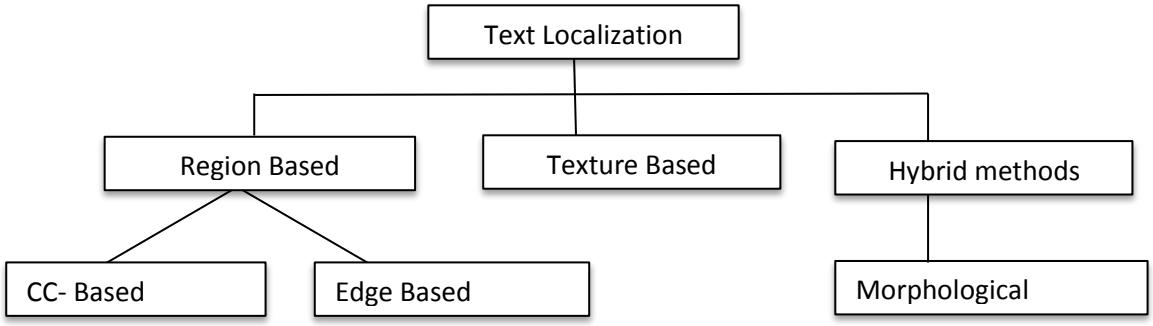


Figure 2.12: Classes of Text Localization Technique

1. Region Based Approach/ Sliding window-based methods

As noted by Sawaki, Murase, & Hagita, (2000); Yadav & Kumar, (2016), the region based approach uses the properties of the color or gray scale in the text region or their differences to determine the corresponding properties of the background. This method groups pixels that displays similar qualities together (Chavre, 2015) ; there is also an ample distinction between the text and the background. Sub structures are identified and merged to form homogeneous region which are marked with bounding boxes. The method is further divided into two: connected component based and edge based method.

a. **Connected component based method (CC-based)**

This approach uses bottom up approach in which successively large components are formed by grouping small similar components (such as edge, color, stroke width, size and gradient information) of the input images to form a connected components. This process iteratively continues until all the regions of the image are identified. after which the non-text regions of the image are filtered by considering a geometrical analysis or conditional random fields (CRFs) which merges the components using the spatial arrangement of the components and mark the boundaries of the text regions. This approach is widely used because of its low computational cost and the outputs can closely cover text regions (Chavre, 2015).

b. Edge Based Methods: Edges are those portions of the image that corresponds to the object boundaries. There is a major contrast in the text and the background and this drastic variation forms the edges. The non-text region of the images is filtered out by firstly identifying the edges of the text boundaries and by applying various heuristics" (Chavre, 2015;

Yadav & Kumar, 2016). For identification of edge in an image, an edge filter is generally incorporated while an averaging operation or a morphological operator is used for merging stage.

Generally, the region-based approach is associated with the following challenges:

i. not efficient for addressing a great quality of classes,

The confusion of the inter-characters and intra-characters as outlined in Figure 2.13 (Mahmood, 2020).

ii. This technique is also very slow as the image requires processing through multiple scales

When a window is made up parts of two characters that are close to each other, it's appearance might be that of a different character altogether. The window in Figure 2.13 (a) includes sections of characters 'o' which may be consider as 'x'. A section of the character 'B' may be considered as 'E' as shown in Figure 2.13(b)

Figure 2.13: Inter-character and Intra-character confusion

(a) Inter-character confusion: A window covering parts of the two o's is erroneously identified as x.

(b) Intra-character confusion: A window covering a section of the character B is taken as E.

2. Texture Based Approach

This method uses textural features of the text. It is "based on the criteria that text in images is different from background in textural characteristic. These properties distinguished the text from its background. Parameters such as energy, contrast, correlation, and entropy define the texture of an image" (Gao, & Yang, 2001; Su *et al*., 2012; Chauhan *et al.*, 2016). "Image is generally scanned in multi-scales in this technique using sliding window, and then texture analysis approaches, such as discrete cosine transform (DCT), Fourier transform, wavelet coefficients, spatial variance and Gabor filters" (Chavre, 2015; Jung, 2001) are used to obtain texture information which led to its high computational cost. Though, this method is much stronger than the CC methods, they only deal with horizontal texts, very sensitive to scale change and rotation and also very Complex to handle (Mahmood, 2020).

3. Hybrid / Mathematical Morphological Approach

The morphological approach is concerned with processing images based on their shapes by dealing with them as sets of points. This approach was proposed basically to enhance detection of video texts under complex backgrounds. Morphological Approach "is a topological and geometrical based approach for image analysis. This approach is used for character recognition and document analysis since operations like translation; rotation and scaling do not have any effect on the geometrical shape of the image. This method works robustly under different image alterations and is used to describe the distribution of stroke inside each over-segmented piece where stroke unit is represented as a text-like segment inside a piece" (Mahmood, 2020).

4. Maximally Stable Extremal Regions (MSER)

MSER (maximally stable extremal regions) is employed within computer vision as a technique of blob detection in images. This technique was developed to establish relationships among elements of images with varying perspectives. "MSER follows a region growing approach, but rather than using the distance to edges, it uses image intensities. The idea behind this is to locate regions with even intensity and to maintain them when images are captured with varying illumination conditions or from different viewpoints". As opposed to region-based approach, MSER-based techniques is robust in the employment of the character extraction, even when confronted with low quality images, strong noises, low resolution etc. with an efficient implementation. The MSERs algorithm can detect the majority of characters. As noted by Abraham, (2015) "the MSER algorithm input is a gray scale image. It is binarized with a threshold t iterating from 0 to 255. 0 stands for completely black and 1 stands for completely white. White regions are selected as extremal regions. This can be applied conversely also. It can be found that for how many successive images in the sequence this extremal region stays the same. Select images which are exactly the same in at least R successive images by selecting a threshold value R. Such regions are called Maximally Stable Extremal Regions." MSER tracking "requires a data structure that can be efficiently built and managed. The component tree is a structure which allows the detection of MSERs within an image and, in addition, constitutes the basis for MSER tracking. The component tree is a rooted, connected tree and can be built for any image with pixel values coming from a totally ordered set. Connected region within the input image are represented by each node of the tree. The nodes of the component tree are identified as connected regions within binary threshold images. Notwithstanding,

MSER algorithm is sensitive to blurred image so gradient amplitude is added to it to produce edge-preserving MSER" (Abraham, 2015).

2.5 Scene Text Detection Based on Deep Learning Approaches

Deep Learning Approaches have been shown to handle the challenges associated with scene text more elegantly than the traditional machine learning approach. The reasons for this are outlined next.

a. **Decision Boundary**

A decision boundary is a surface that separates data points belonging to different class labels. Decision Boundaries are not only confined to just the provided data, they span through the entire feature space that is trained on. The model can predict a value for any possible combination of inputs in our feature space. If the training data on is not varied enough, the overall topology of the model will generalize poorly to new instances. So, it is important to analyze all the models which can be best suitable for 'varied dataset, before using the model into production. Decision boundary is therefore used to determine if a given data point belongs to a positive class or a negative class. It also helps to learn how the selected training data affects performance and the ability for the model to generalize.(Suchismita, 2021).

b. **Feature Engineering**

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. This is achieved using feature extraction and feature selection. Feature extraction extracts all the required features for the problem statement. While feature selection select the

important features that improve the performance of the machine learning or deep learning model.

Research has shown that extracting features manually from an image can be very challenging and time consuming because of the need for a very strong knowledge of the subject as well as the domain.

Though, there are an handful of traditional approaches to scene text detection as noted by (Nixon & Aguado, 2020; Abraham, 2015; Mahmood, 2020) yet scene text detection based on deep learning such as EAST (Efficient accurate scene text detector) and CRAFT (Character Region Awareness For Text detection) have shown undeniable improvement in the area of performance recently (Baek *et al.*, 2019).

The deep learning methods mostly train their networks to localize word level bounding boxes. Nonetheless, when confronted with difficult cases like curved text, very long and degraded or deformed text, which are hard to detect with a single bounding box, they tend to malfunction. This downside is curbed by character-level awareness which has many advantages when handling challenging texts by linking the successive characters in a bottom-up manner. Regrettably also, character level annotations are not provided by most existing dataset and obtaining character level ground truth becomes very expensive. Hence, Character-Region Awareness for Text detection (CRAFT) for text detection is adopted for the purpose of this research.

2.5.1 CRAFT (Character-Region Awareness for Text Detection)

To effectively detect text areas by exploring each character and affinity between characters, CRAFT which adopts a fully convolutional neural network with VGG-16 as its backbone was designed. VGG 16 architecture is a feature extracting architecture that is used to encode the network's input into a certain feature

representation. The major target of CRAFT is to localize the individual character regions in an image and the localized characters are linked to a text instance. CRAFT predicts two scores for each character: Region Score and Affinity score: the Region Score localizes individual character in the image and the Affinity Score groups each character into a single instance. Afterwards, the affinity and region scores are combined to give the bounding box for each word in the image. Its final output produces two channels as Score Maps: Region Level Map and Affinity Map. The region map indicates the areas where characters are present. The Affinity Map is a pictorial representation of related characters. Red symbolizes the characters have a high affinity and must be merged into a word. Thereafter, the affinity and region scores are combined to give the bounding box of each word. The coordinates followed the order: (left-top), (right-top) (right-bottom), (left-bottom), in (x, y) pair (Ti, 2022). Figure 2.14 shows CRAFT architecture as documented by Baek *et al.*, (2019).

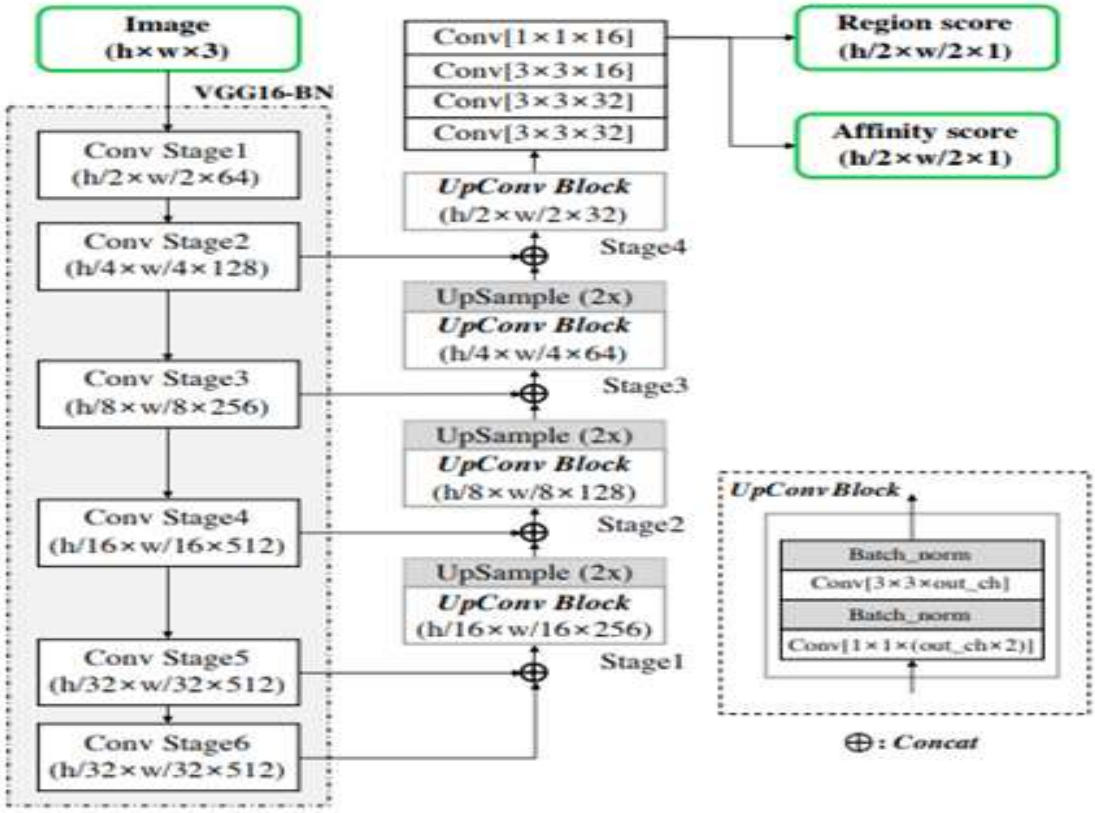


Figure 2.14: General CRAFT Architecture. *Source*: Baek *et al.*, (2019)

2.6 Text Recognition

The next task after text localization and enhancement is text recognition. The task of text recognition is to identify the class of each character in the text. It points out the exact alphabet or number that was detected. This task becomes very easy when dealing with standard print document with regular fonts such as Calibri which can be easily handled by traditional character/word recognition engine. On the contrary, recognition becomes very difficult when the target is not conventional paper document. In captured scene images, more difficult characters are to be recognized such as "calligraphic characters, pictorial characters, multi-colored characters, textured characters, transparent characters, faded characters, touching characters, broken characters, very-large (or small) characters" are also encountered. In addition, handwritten characters are often encountered on notebook and whiteboard, and "handwritten fonts" on signboard and poster are encountered as well (Uchida, 2020). Therefore, to develop a robust text recognition system that will be robust enough to take care of the variations in scene text, the CRNN which leverages the functionalities of CNN, RNN and CTC will be adopted. These approaches are outlined next.

2.6.1 Types of Neural Networks in Deep Learning

Outlined next are the different types of neural networks in deep learning.

a. **Artificial Neural Network (ANN)**

ANN, is a group of multiple perceptrons/ neurons at each layer. ANN is also known as a Feed-Forward Neural network because inputs are processed only in the forward direction. Neural networks are generally organized into multiple layers. Figure 2.15 consists of 3 layers:

.

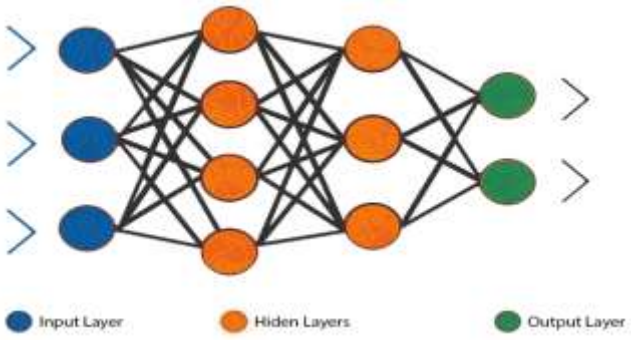


Figure 2.15: Artificial Neural Network

Input layer which accepts the inputs, hidden layer processes the inputs and Output layer that produces the final result. Each of these layers tries to learn certain weights

ANN are generally use to solve problems that are related to Tabular data, Image data and Text data. They have the capability of reading Nonlinear functions, weight that map any input to output and complex relationship between input and output. Hence, they are referred to as Universal Function Approximators. ANN loses the spatial features (the arrangement of the pixels in an image) of an image. it is also not capable of capturing sequential information in the input data which is required for dealing with sequence data.

b. Recurrent Neural Network (RNN)

RNN has a recurrent connection on the hidden state of ANN. Here, the output of the processing nodes is saved and the result is fed back into the model. This implies that the information is not passed only in one direction. Hence, the model is said to learn to predict the outcome of a layer. In RNN model, each node acts as a memory cell,

continuing the computation and implementation of operations. When the network predicts wrongly, the system self-learns and continues working towards the correct prediction during back propagation. RNN can widely be applied in Time Series data, Text data and Audio data. It has the capacity to remember information through time. They are used with the convolutional layers to extend the effective pixels neighborhood. Though, RNN is very difficult to train. Also it has vanishing and exploding gradient problem

c. Convolution Neural Network (CNN)

CNN is a type of neural network that uses the operation of convolution to extract relevant features from an image. It has a wide range of application which is prevalent in image and video processing. It uses a variation of multilayer perceptrons and contains one or more convolutional layers that can be either entirely connected or pooled. The building blocks of CNNs are filters and kernels precisely. The relevant features from the input are extracted with Kernels using the convolution operation. As noted by research, CNN has a very high accuracy in image recognition problems. It detects the important features automatically without any human supervision. Also, it can be modified in the training period to enable extracting most relevant features. Contrarily CNN does not encode the position and orientation of object. It lacks the ability to be spatially invariant to the input data and also requires Lots of training data.

d. Connectionist Temporal Classification (CTC)

CTC is a type of Neural Network output use for tackling sequence problems like handwriting and speech recognition where the timing varies. Using CTC ensures that aligned dataset are avoided. This makes the training process more straightforward. “CTC helps in calculating the loss while training the neural network and decoding the output of the neural network. CTC model is used to

efficiently evaluate all possible alignment of hand gesture through dynamic programming and check consistency via frame-to-frame for the visual similarity of hand gesture in the unsegmented input stream" (Patel & Makwana, 2020).

5. Convolutional Recurrent Neural Network (CRNN)

The problems associated with optical character recognition are mostly image-based sequence recognition. And the most befitting neural networks for sequence recognition problem are Recurrent Neural Networks (RNN) whereas Convolution Neural Networks (CNN) are best for an image-based problem. Therefore, to address the OCR challenges to a reasonable extent, there is need to leverage the functionalities of both RNN and CNN. Hence, CRNN is a combination of CNN, RNN, and CTC loss for image-based sequence recognition tasks, such as scene text recognition and OCR. The architecture of the convolutional layers is based on the VGG-Very Deep architectures. The architecture consists of convolutional layers which extract feature maps automatically and a recurrent network used for predicting the text from each frame of the feature map sequence. The per-frame prediction from the recurrent network is translated into a label sequence using a transcription layer. Finally, the transcription layer used Connectionist Temporal Classification (CTC) loss function to predict the output at each point. Figure 2.16 Shows the CRNN architecture as documented by Kambala et al., (2020). The architecture is the combination of CNN and RNN along with transcription layers and Connectionist Temporal Classification (CTC). The CRNN architecture does not require character segmentation. Feature maps are extracted by the convolutional part of the neural network from the input region that contains the detected text region. The deep bidirectional recurrent neural network following the CNN predicts the encoded label sequence which when decoded by CTC gives the final output.

This simply implies that the CRNN network implementation as noted by (TheAILearner, 2022) can be split into the following steps:

1. Collecting Dataset
2. Preprocessing Data
3. Creating Network Architecture
4. Defining Loss Function
5. Training Model
6. Decoding Outputs from Prediction

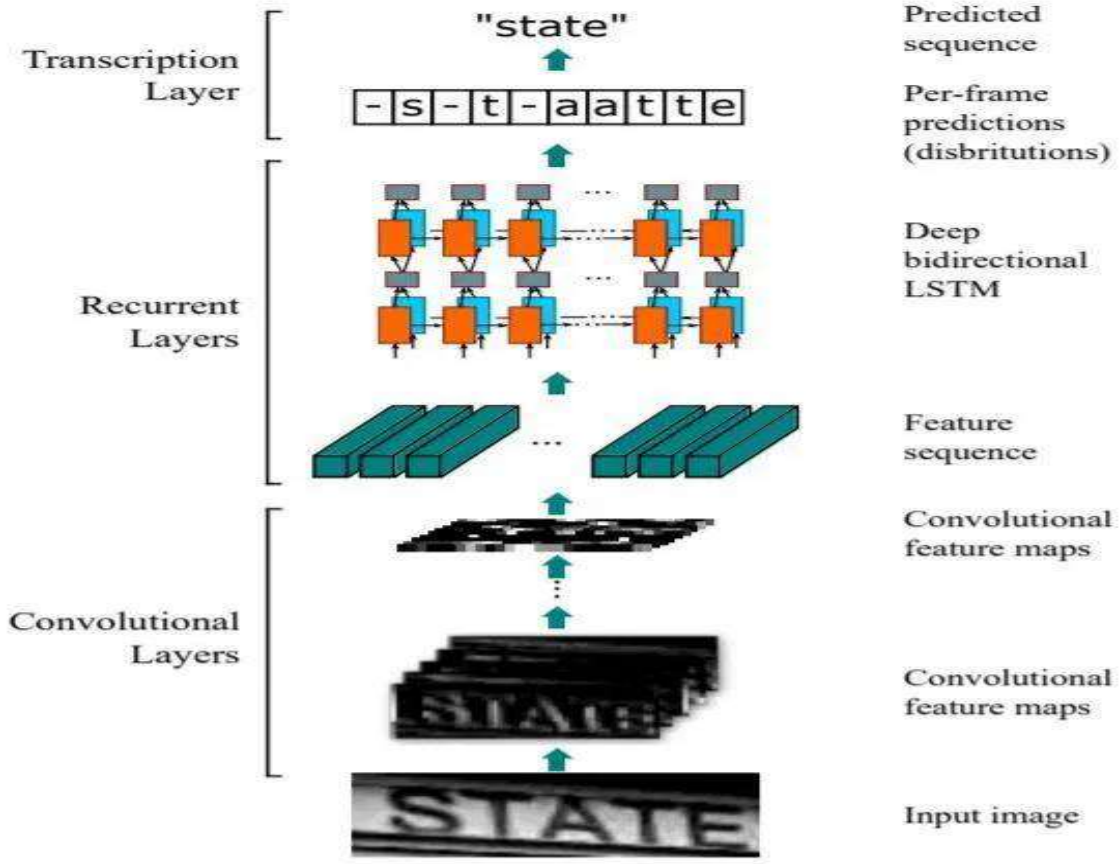


Figure 2.16: The CRNN Architecture

2.7 Approaches to Camera-Based Text Detection and Recognition

This section reviews research works that are closely related to this study. This include the review of the different approaches, methods and algorithms for text localization and feature extraction in scene images, complex background and document with multipart.

2.7.1 Locality Preserving Projection (LPP)

LPP is one of the prevailing feature extraction techniques as noted by (Li *et al.*,2008). It is a dimensionality reduction Technique whose main idea is described as follows: given a set of N-dimensional data $\{x_1, x_2, \ldots, x_n\}$ in R^N from the L classes, LPP aims to find a transformation matrix W to map these points to a set of points $\{z_1, z_2, \ldots, z_n\}$ in R^d (d _ N) based on one objection function z_i (i = 1, 2, . . ., n)" which is given in equation 2 (Li *et al.*,2008).

$$\min \sum_{i,j} \|Zi - Zj\|^2 Sij, \qquad (2)$$

Where: S is a similarity matrix with weights characterizing the likelihood of two points, and zi is the one dimensional representation of x_i with the projection vector w, i.e., $z_i = w_T x_i$. By minimizing the objective function, LPP encounters a major drawback when the neighboring mapped point's z_i and z_j are far apart. Hence, information in the original space is preserved after dimension reduction with LPP by keeping the mapped points close where their original points are close to each other. That notwithstanding, LPP is limited by its ignorance of label information when applied to the classification task. Therefore, as a way of improving on LLP techniques, (Li *et al.*, 2008) proposes a novel local structure based feature extraction method, called class-wise locality preserving projection (CLPP). This method uses class information as a guide to procedure of feature extraction. Also,

the local information and the class information are used to retrieve the local structure of the original data considering certain kind of similarities between data point. To enhance the performance of CLPP on a nonlinear feature extraction, they applied kernel trick to the proposed class wise locality preservation projection. This new version called Kernel CLPP algorithm was evaluated on ORL face database and YALE face database.

2.7.2 Edge Detection Techniques

Edge Detection Technique is another approach used for feature extraction for a Camera-based OCR. "An edge may be viewed as a set of connected pixels that forms a boundary between two disjoints regions. It is basically used for detecting and extracting image features. It identifies points in a digital image where brightness of images changes sharply and finds discontinuities". (Kumar & Saxena, 2013; Hinde & Elfkihi, 2010). It is used to detect all edges in the image and as well preserve the most important structural features of that image by finding sharp contrasts. Hinde A., Elfkihi S., (2010) applied "this technique in their proposed system for text detection in images which is based on texture features and connected components analysis. This technique as discussed below, exploits various differential operators to detect changes in the gradients of the grey". The two main operators used by edge detection technique are:

1. First Order Edge Detection Or Gradient Based Edge Operator
2. Second Order Edge Detector/ Laplacian Based Operator.

a. First Order Edge Detection or Gradient Based Edge Operator

This approach is based on the use of a first order derivative, or gradient based. If the point image is given as I(i , j), then image gradient is given by equation 3.

$$\nabla I(i,j) = i\frac{\partial I(i,j)}{\partial i} + j\frac{\partial I(i,j)}{\partial j} \qquad (3)$$

Where:

$\frac{\partial I(i,j)}{\partial i}$ is the gradient in the i direction

$\frac{\partial I(i,j)}{\partial i}$ is the gradient in the j direction

The gradient magnitude can be computed by the equation 4 (M. Kumar & Saxena, 2013)

$$\|G\| = \sqrt{\left(\frac{\partial I}{\partial i}\right)^2 + \left(\frac{\partial I}{\partial j}\right)^2} \quad \text{or} \quad |G| = \sqrt{Gi^2 + Gj^2} \quad \text{or} \quad \theta = \arctan\left(Gj/Gi\right). \qquad (4)$$

Note: the gradient magnitude is used to obtain the edge strength and the gradient direction is always perpendicular to the direction of edge.

The operators used in first order edge detection techniques are further subdivided into:

i. Classical Operators (includes Robert, Sober , Prewitt)
ii. Canny Edge Detector

i. Classical Operators

The operators involved in classical approach are outlined next.

a. **Robert Operator:** this operator measures the spatial gradient of an image. It takes input image as gray scale image and produces edges that are contained in that image. The advantage of this operator is its simplicity, but it is not symmetric and also, it is unable to detect edges that are multiplies of 45 degrees and having small kernel makes it highly sensitive to noise and as such, it is not compatible with the recent technology. (Das, 2016; M. Kumar & Saxena, 2013;

Shrivakshan, 2012). It uses two kernels of 2 x 2 dimensions as shown in equation 5.

$$Dx = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \text{ and } \quad Dy = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{5}$$

b. **Sober Operator:** is a "discrete differentiation operator used to compute an approximation of the gradient of image intensity function for edge detection. It gives a corresponding gradient vector or normal vector at each pixel of an image. This technique has better noise suppression characteristics when compared with Prewitt kernel". When compared with Robert operator, "it has low computation ability but less sensitive to noise because of its high kernel. It applies two dimensional spatial gradient measurements on an image and also highlight spatial component that belongs to edge" (Agarwal, 2015; Kumar & Saxena, 2013). Equation 6 shows the 3 x 3 two kernels used by Sober Operator.

$$Di = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +2 \end{bmatrix} \text{ and } Dj = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \tag{6}$$

c. **Prewitt Operator:** this operator is suitable for measuring the magnitude and orientation of the edges. This technique uses a pair of 3x3 convolution masks of 8 directions, and it evaluates the edge directions directly with the maximum response from the mask. It is very closely related to sober operator but has different Kernel and better performance than Sober. (Das, 2016; Kumar & Saxena, 2013). Masks used for Prewitt Edge Detection Technique is shown in equation 7. Figure 2.17 shows the flow chat of general algorithm for classical operators used by first order edge detection (Das, 2016; Kumar & Saxena, 2013).

$$Di = \begin{bmatrix} -1 & 0 & +1 \\ -1 & 0 & +1 \\ -1 & 0 & +1 \end{bmatrix} \text{ and } Dj = \begin{bmatrix} +1 & +1 & +1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix} \quad (7)$$

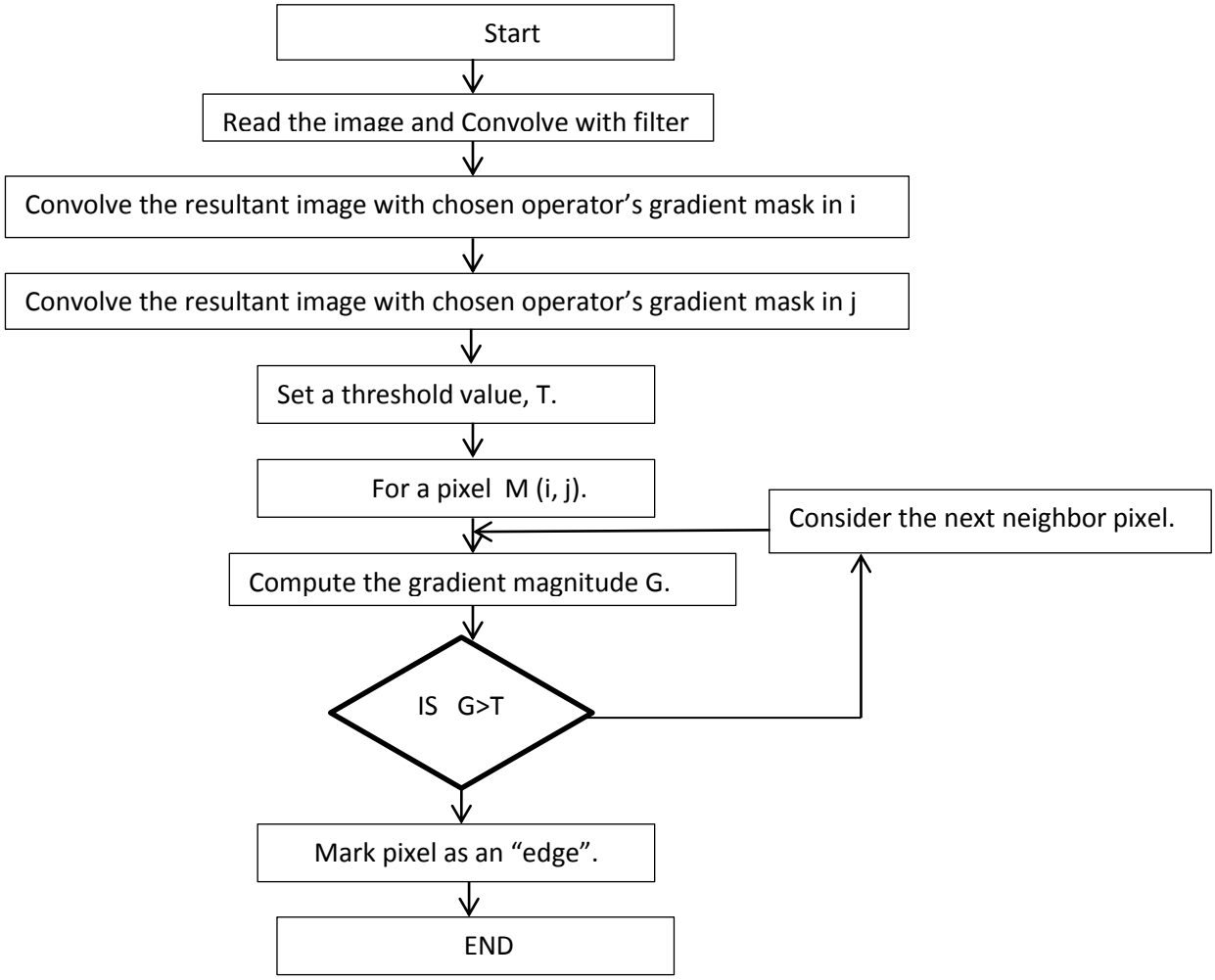


Figure 2.17: General Algorithm for Classical Operators

ii. Canny Edge Detector

Canny Edge operator is considered a standard and mostly used edge detection approach than other existing methods. It separate noise from the image before extracting edges. Its algorithm follows the steps given below:

Step1: Image smoothing: here, the image is blurred to remove noise using the Gaussian filter: $F(i,j) = G*I(i,j)$ (**8**)

Step2: Find gradient: Compute the gradient magnitude and orientation of the smoothed image using finite-difference approximations for the partial derivatives or other classic operators. The 3*3 matrices shown in equation 2.9 shows Sobel operator used to determine the gradient at each pixel of smoothened image.

$$Di = \begin{bmatrix} -1 & 0 & +1 \\ 2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \quad \text{and} \quad Dj = \begin{bmatrix} +1 & +2 & -1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{9}$$

*Gi = Di * F(i, j) And Gj = Dj * F(i, j) (gradients in i and j directions)*

Therefore edge strength or magnitude of gradient of a pixel is given by equation 10 and equation 11 shows the direction of gradient.

$$G = \sqrt{Gi^2 + Gj^2} \tag{10}$$

$$\theta = \arctan\left(\frac{Gj}{Gi}\right) \tag{11}$$

Where G_i *and* G_j *are the gradients in the i- and j-directions respectively.*

Step3: non-maxima suppression: is carried out to preserves all local maxima in the gradient image, and deleting everything else, the blurred images are also sharpened; this result in thin edges.

Step4: Hysteresis thresholding/ Double thresholding: here, "noise generated in the non-maxima suppression is removed. Two thresholds are used by the Canny detector—t_{high} , and a t_{low} threshold. Therefore, edge pixels stronger than the high threshold are labeled as "strong", edge pixels weaker than the low threshold are suppressed, and edge pixels between the two thresholds are labelled as weak"

(Agarwal, 2015; Das, 2016; Kumar & Saxena, 2013; Shrivakshan, 2012 ; Khan *et al.*, 2019).

Eg "For a pixel M (i, j) having gradient magnitude G following conditions exists to detect pixel as edge:

If $G < t_{\text{low}}$ then discard the edge.

If $G > t_{high}$ *then keep the edge.*

If $t_{low} < G <$ *thigh and any of its neighbors in a 3 ×3 region around it have gradient magnitudes greater than* t_{high}*, keep the edge.*

If none of pixel (x, y)'s neighbors have high gradient magnitudes but at least one falls between t_{low} and, t_{high} search the 3 × 3 region to see if any of these pixels have a magnitude greater than t_{high}. If so, keep the edge.

Else, discard the edge".

Step5: Edge tracking: in this step, the strong edges are immediately included in the final edge image. Only weak edges connected to strong edges are included.

Canny Edge Detection Technique Masks uses 3*3 matrices. The flowchart is as shown in Figure 2.18 (Kumar & Saxena, 2013).

$$Gx = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad Gy = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{12}$$

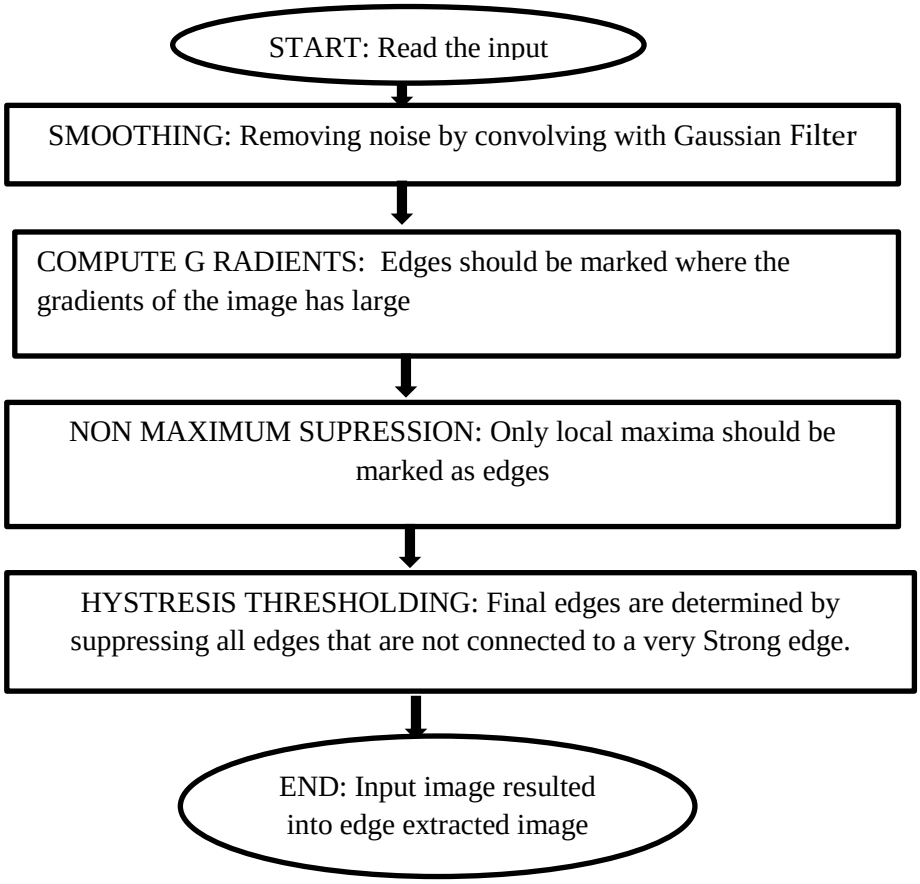


Figure 2.18: Flow chart of canny edge detection algorithm

b. Second Order Edge Detector/Laplacian Based Operator

In second order edge detector also known as Laplacian Based Operator, "a pixel is marked as an edge at a position where second derivative of an image becomes zero". The Laplacian operator of a second order derivative for a 2D image I (i, j) is given by equation 13 (Kumar & Saxena, 2013).

$$\nabla^2 = I(i,j) = \frac{\partial^2}{\partial X^2} I(i,j) + \frac{\partial^2}{\partial y^2} I(i,j) \qquad (13)$$

Laplacian of Gaussian as noted by (Kumar & Saxena, 2013; Kurchaniya & Dixit, 2017) is "a gradient based operator which uses the Laplacian to take the second derivative of an image. it locates edges then search the zero crossing between the double edges. It exploits Gaussian operator to reduce noise while the laplacian operator is use for sharp edges detection. It contains the pair of 3x3 convolution mask". This technique is based on second order derivative as shown in equation 14 shows the Masks used for LOG Edge Detection Technique.

$$Gx = \begin{bmatrix} +1 & +1 & +1 \\ +1 & -8 & +1 \\ +1 & +1 & +1 \end{bmatrix} \quad , \quad Gy = \begin{bmatrix} -1 & +2 & -1 \\ +2 & -4 & +2 \\ -1 & +2 & -1 \end{bmatrix} \tag{14}$$

However, the Laplacian of Gaussian (LoG) as documented by Kumar & Saxena, (2013) suffers from two main limitations. These are outlined next.

1. False edges which are generated responses that do not correspond to edges,

2. Localization error may be severe at curved edges. (Kumar & Saxena, 2013)

Flow chart of general algorithm for Laplacian of Gaussian operator is shown in figure 2.19

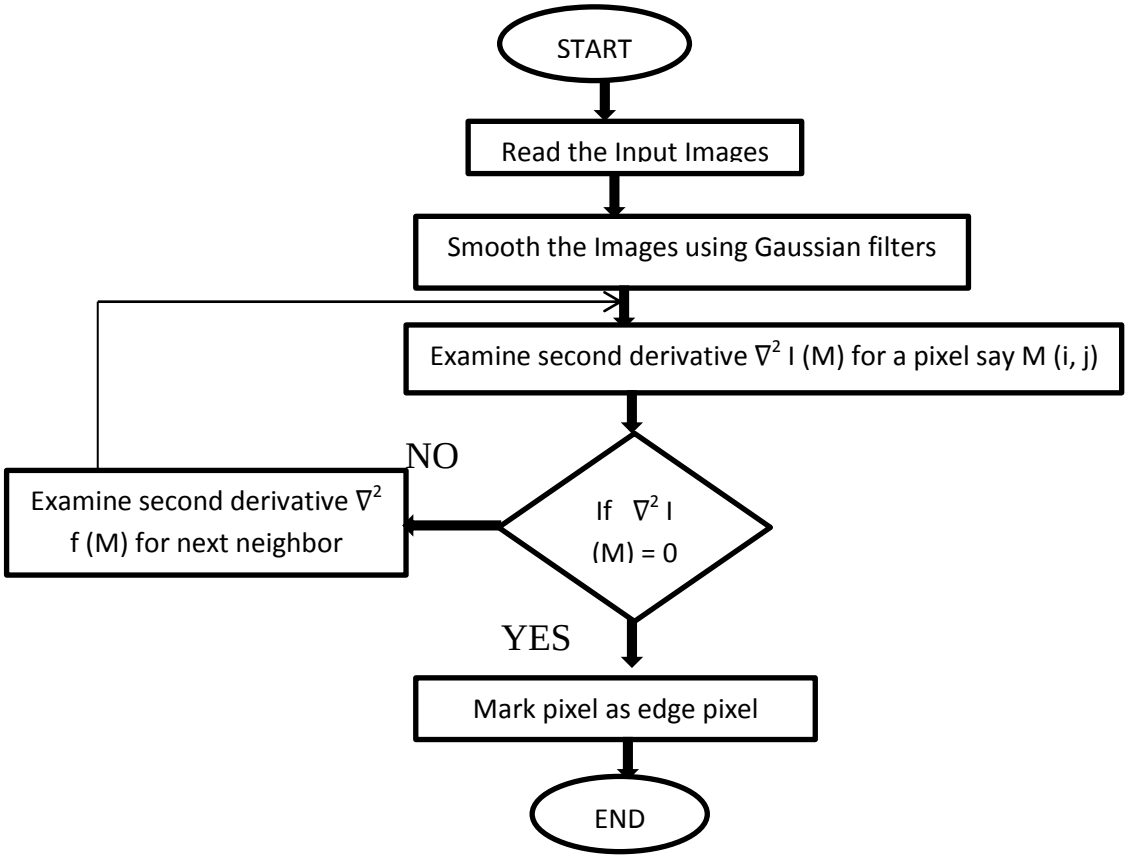


Figure 2.19: Flow Chart of General Algorithm for Laplacian of Gaussian Operator.

Research has shown that canny edge detection gives a state of the art performance (Das, 2016; Kumar & Saxena, 2013; Kurchaniya & Dixit, 2017). This is attributed to the following points:

1. Adaptive in nature: unlike classical operators that cannot be adapted to a given image because they have fixed kernel, the performance of canny algorithm depends on adjustable variables or parameters like Sigma. The larger the value of sigma, the more blur the image appear; this can be revised by choosing a smaller value of sigma.

2. Tolerates Noise more than the classical operators. Canny method of edge detection is less sensitive to noise. It uses Gaussian filter which removes noise at a great extent.
3. Removes streaking Problems by applying hysteresis techniques which uses two threshold values t_{low} and t_{high} . Unlike the classical operators that uses single threshold and thus finds it difficult to handle streaking.
4. Enhanced Localization: Canny operator when compared to LoG operator provides edge gradient orientation which enhances good localization while LoG finds it difficult to locate edge orientation.

2.8 Mobility aid for the Visually Impaired

Enabling the vision impaired to live independently both on mobility and accessing print text is an area of research that cannot be overlooked nor over-emphasized. This section explores different works done by different novel researchers on enabling the Vision Impaired persons to work independently on indoors, outdoors and even both open and enclosed arrears, as to ascertain how well the existing systems are working to ensuring a comfortable and independent living for people with vision impairment. It also explores the methods used, how efficient these methods are and the areas that can be improved on as to come up with a more suitable approach to developing a world class system for the concerned community.

2.8.1 Outdoor Navigation Systems for the Vision Impaired

Navigation in an open space is a very challenging and difficult task for the visually impaired persons. As a way of solving this problem, Borenstein & Ulrich, (1997) developed a computerized travel aid for the active guidance of blind pedestrians. Their paper introduces “the GuideCane, a novel device designed to help blind or

visually impaired travelers to navigate safely and quickly among obstacles and other hazards faced by blind pedestrians". To further improve on the work done by (Borenstein & Ulrich, 1997), Ercoli *et al*., (2013) carried out a research on "A wearable multipoint ultrasonic travel aids for visually impaired". This study was based on the findings on the limitations of the use of white cane for mobility aid for the visually impaired. However, "future work should be geared towards improving the transmission of the information to the VI user by vibro-tactile interfaces and allowing the use of the proposed system in conjunction with the white can without interferences" as implemented by (Emily *et al.*, 2014; Bhatlawande *et al*., 2014; Siva *et al.,* 2019; Aiswarya & Shiji, 2019).

Still on the quest to improve the mobility and navigation of the blind community, (Bousbia-salah & Fezari, 2007; Chukwunazo & Onengiye, 2016), carried out a study on "A Navigation Tool for Blind People". The paper presented "describes the development of a navigation aid in order to assist blind and visually impaired people to navigate easily, safely and to detect any obstacles. The system is based on a microcontroller with synthetic speech output". However, since it is difficult to ascertain the location of the PVI globally with the new system, it is then required that future development should incorporate global positioning system in order to get the user position information. On that basis, (Morad, 2010; Gawari & Bakuli 2014; Vel *et al.*, 2018), carried out a study on "GPS Talking for Blind People". In this study, they designed a device to help the blind people to navigate the environment without asking external help. Conversely, the system may not function effectively indoors because of the GPS which works efficiently in a very wide and open place.

Dakopoulos & Bourbakis, (2010) worked on "Wearable Obstacle Avoidance Electronic Travel Aids for Blind: A Survey". There are three main categories of

portable or wearable navigation systems. These systems include Electronic travel aids (ETAs), electronic orientation aids (EOAs), and position locator devices (PLDs). This paper however, present "a comparative survey among portable/wearable obstacle detection/avoidance systems (a subcategory of ETAs) in an effort to inform the research community and users about the capabilities of these systems and about the progress in assistive technology for visually impaired people." However, having carefully examined the existing ETAs, the research made the following recommendations on future systems characteristics: "1. since the user will still hold the white cane which is almost inevitable, the system has to be Free-hands i.e. not requiring from the user to hold them. "2) Free-ears: despite the advantages of echolocation, spatial sound, and similar techniques, the user's ability to listen environmental should not be interfered; 3) Wearable: it offers flexibility to the users and utilizes the advantages of wearable technologies; 4) Simple: easy to use (operation and interface not loaded with unnecessary features) and without the need of extensive training period".

Muhammad, (2012) designed a "Smartphone based Guidance System for Visually Impaired Person". The system is aimed at facilitating the mobility of visually impaired person. However, this system uses the edges from different surrounding structures and it loses its accuracy once these edges are not present. Therefore, Future work may look into paths which are not straight, in more complex environments and with a lot of variations in occlusion. Fawaz *et al*., (2013) also, proposed a "Smart and Intelligent System to assist the blind community based on advance Technologies." Still on mobility of People with visual impairment, Brock, (2013) carried out a research on "Interactive Maps for Visually Impaired People: Design, Usability and Spatial Cognition". For visually impaired people, knowing the geography of an urban environment is fundamental. Bharambe *et al.,* (2013)

presents an affordable and a supportive device for visually impaired people that can be used as an artificial eye.

Yet on the journey of improving the life of PVI, (Joseph *et al.*,2015) carried out a study on "Being Aware of the World: Toward Using Social Media to Support the Blind with Navigation". The system "lays the ground work for assistive navigation using wearable sensors and social sensors to foster situational awareness for the blind. Sohl-dickstein *et al*., (2015) present a prototype assistive device called (Sonic eye) to aid in navigation and object perception based on ultrasonic echolocation and spatial hearing principles that aid visually impaired people for their self-regulating mobility, in their research work titled "A device for human ultrasonic echolocation".

Azlina Aziz, (2016) argued that "having a sight impairment should not limit one's opportunity to be socially included and obtain the many benefits of being in a green space. Following this argument, he carried out a research on the value of green space to people with a late onset visual impairment". On that note, Zhou *et al.*, (2016) developed "A Smart \"Virtual Eye\" Mobile System for the Visually Impaired" to enhance the independent living of the visually impaired people by providing them with information on their surroundings. Aladr´en *et al.*, (2016) as well worked on "Navigation Assistance for the Visually Impaired Using RGB-D Sensor with Range Expansion". The proposed system called "(NAVI) is able to assist or guide people with vision loss, ranging from partially sighted to totally blind, by means of sound commands provided by an RGB-D camera that collects visual and range information from the surroundings".

Eunjeong & Eun, (2017); Nair & Tsangouri, (2018); Yadav *et al.,* (2018); Ullah *et al*., (2018), having considered way finding as one of the major challenges confronting the VIP, worked on "A Vision-Based Wayfinding System for Visually

Impaired People Using Situation Awareness and Activity-Based Instructions, that can automatically recognize the situation and scene objects in real time and help them find their way in an unfamiliar indoor environment". Noman *et al.,* (2017) Design and implemented a microcontroller based assistive robot for person with blind autism and visual impairment. As state by (Mohanty *et al.,* 2017), the most valuable gift we have is the eye and it plays a very crucial role in our day to day activities, of which life is unbearable without it. Based on this, these researchers proposed an android based navigational shoes system titled "Smart Shoe for Visually Impaired" that will enable the visually impaired move comfortably in an unfamiliar public places. Yeonju, *et al.,* (2017) supported the above concept by proposing a novel navigation system that assists the visually impaired to travel in unfamiliar indoor environments independently in their research work titled "Indoor Navigation Aid System Using No Positioning Technique for Visually Impaired People". Ramadhan, (2018) presents a "wearable smart system to help visually impaired persons (VIPs) walk by themselves through the streets, navigate in public places, and seek assistance".

Having seen the challenges experienced by visually impaired visitors while visiting museum's exhibition, (Vaz *et al.*, 2018) carried out a research on the topic "Designing an Interactive Exhibitor for Assisting Blind and Visually Impaired Visitors in Tactile Exploration of Original Museum Pieces". In the research titled "Assistive device for orientation and mobility of the visually impaired based on millimeter wave radar technology - Clinical investigation results", (Kiuru *et al.,* 2018) present clinical investigation results of a novel, wearable, orientation and mobility assistive device that uses radar technology.

One of the major challenges of the visually impaired persons is moving smoothly in an open space. On that note, (Islam & Sadi, 2018) went into a study on "Path Hole

Detection to Assist the Visually Impaired People in Navigation". The researchers, having seen path hole detection a prominent issue to aid the visually impaired people; proposed a solution by detecting path hole on the road surfaces using Convolution Neural Network. "The proposed system is able to classify the road surfaces with path hole and non-path hole. Cardillo *et al.*, (2018) worked on "An Electromagnetic Sensor Prototype to Assist Visually Impaired and Blind People in Autonomous Walking". Compared to the already existing Electronic Travel Aids devices, the proposed system exhibits better performance, noise tolerance and reduced dimensions". However, to improve on the proposed system, the suitability of the system needs to be improved by integrating all the subunits over the same circuit board. (Elmannai & Elleithy, 2018) developed "A Highly Accurate and Reliable Data Fusion Framework for Guiding the Visually Impaired". This research was motivated based on "the fact that 90% of VI people live in developing countries. Several systems were designed to improve the quality of the life of VI people and support their mobility. Unfortunately, none of these systems are considered to be a complete solution for VI people and these systems are very expensive". This paper presents an intelligent framework for supporting VI people.

Mancini *et al.*, (2018) in their research work titled "Mechatronic System to Help Visually Impaired Users during Walking and Running, introduced a monocular vision-based system to aid visually impaired people in walking, running, and jogging in outdoor environment". This was further improved in the works done by (Sotomayor *et al.,* 2019; Sotomayor *et al.,* 2019). Their system aid in detecting obstacles and water puddles on their way.

2.8.2 Indoor Navigation Systems

The rate at which vision loss especially in the underdeveloped countries is escalating is quite disturbing. On this regard, (Mehta *et a*l., 2011) proposed an indoor navigation system in their research titled "VI-Navi: A Novel Indoor Navigation System for Visually Impaired People". However, the developed system is restricted to enclosed surroundings which imply that the system may not function well in an open environment. Their work were further supported by the research done by (Guerrero *et al*., 2012; Nakajima & Haruyama, 2013; Legge *et al*., 2013; Al-Shehabi *et al*., 2014). (Wang & Tian, 2012) worked on the topic "Detecting stairs and pedestrian crosswalks for the blind by RGBD camera". The proposed system is geared towards improving the safe navigation of persons with visual impairment in unfamiliar places.

Mohamad *et al*., (2013), still on the quest to improving the life of people with visual impairment, carried out a research on "Smart walking stick - an electronic approach to assist visually disabled persons. the main purpose of their paper is "based on abating the disabilities of blindness by constructing a microcontroller based automated hardware that can corroborate a blind to detect obstacles in front of him/her instantly".

For wearable devices, (Zhigang *et al*., 2014) designed a "Wearable Navigation Assistance For The Vision-Impaired". The assistive device includes "a sensor that detects information using a first modality; an actuator that conveys information using a second, different modality; and a controller that automatically receives information from the sensor and operates the actuator to provide a corresponding actuation". Martinez-sala *et al.,* (2015) in their work titled "Design, Implementation and Evaluation of an Indoor Navigation System for Visually Impaired People

developed Universal Advanced Guidance System in Enclosed Areas (SUGAR)", that provides high accuracy in terms of location.

(Romić *et al.*, 2015; Khaliluzzaman *et al.*, 2016; Alam *et al*., 2018) carried out a research on the topic "Technology assisting the blind — video processing based staircase detection". This study is aimed at assisting the blind to detect stairs in an unfamiliar environment.

For a system that will aid a visually impaired in both navigation and mobility, (Martin, 2016; Salam & Korial, 2016); Sato & Kitani, 2017)) in their research on "Technology Based Aid for the Visually Impaired", presents a system that can assist a person with a visual impairment in both navigation and mobility. Sarojini *et al.*, (2017) also proposed to develop a "Smart Electronic Gadget for Visually Impaired People" for obstacle detection in the path of visually impaired people. The system is a wearable device which can be fitted in our belt ID tag or hat that has infrared sensors and raspberry pi installed on it. Govardhan *et al.,* (2017) designed a "Smart Object Detector for Visually Impaired". The main aim of this research work is to design and develop a low cost machine to the visually impaired for detecting objects. Kamal *et al.,* (2017) went into a study on the topic "Towards Developing Walking Assistants for the Visually Impaired People". The target of their study is to provide a solution that helps blind to avoid obstacles around them without holding any sticks or other heavy things.

On an attempt to providing Portable mobility device, Bai *et al.,* (2017) worked on "Smart Guiding Glasses for Visually Impaired People in Indoor Environment". The developed system is a smart guiding tool similar to an eyeglass, which aid the visually impaired people in moving safely. In with bai et al., (Cheraghi *et al.,* 2017) carried out a research on "GuideBeacon: Beacon-based indoor wayfinding for the blind, visually impaired, and disoriented." The system was developed to assist individuals with visual impairment with their indoor wayfindings and navigation

within interior surroundings. Diwakara & Surendran, (2018) also carried out a research on "Real Time Indoor Path Navigation Wearable for Visually Impaired". Having considered the challenges and difficulties encountered by the visually impaired in moving from one fixed point to another, they went into a research as to proffer a lasting solution to the existing challenges.

2.8.3 Indoor and Out Door Mobility Aid

On ensuring both indoor and outdoor movement for the visually impaired, (Kammoun *et al*., 2012) designed "Navigation and space perception assistance for the visually impaired: The NAVIG project". The system is aimed at improving the mobility and easy navigation of people with visual impairment. Tapu *et al*., (2013) went into a research on "A Smartphone-Based Obstacle Detection and Classification System". In this study, "they introduced a real-time obstacle detection and classification system designed to assist visually impaired people to navigate safely, in indoor and outdoor environments, by handling a smartphone device for Assisting Visually Impaired People". Alghamdi *et al., (*2014) worked on "Accurate Positioning Using Long Range Active RFID Technology to Assist Visually Impaired People" to aid visually impaired people both in their indoor and outdoor navigation.

Having measured the number of challenges faced by the blind and the virtually impaired person in their day to day life, (Barati & Delavar, 2015) went into a research on "Design And Development Of A Mobile Sensor Based The Blind Assistance Wayfinding System". Cecílio *et al.,* (2015) worked on Information Tracking System for Real-time Guiding of Blind People. The researchers noted that among the activities affected by visual impairment, navigation plays a fundamental role, since it enables the person to independently move in safety". Based on this,

they proposed “an innovative navigation and information system to help the navigation of blind people within new environments” (e.g. shopping center, public office building).

Yang *et al*., (2016) worked on “Expanding the Detection of Traversable Area with RealSense for the Visually Impaired”. The researchers “proposed an effective approach to expand the detection of traversable area based on a RGB-D sensor, the Intel RealSense R200, IR camera, RGB camera, processor, attitude sensor and bone conducting headphone”. This is compatible with both indoor and outdoor environments.

Wafa & Elleithy, (2017) in their quest on how to improve the quality of life of the VIPs, conducted a research on “Sensor-Based Assistive Devices for Visually-Impaired People: Current Status, Challenges, and Future Directions”. Having considered the limitations in the capabilities of the existing system, present “a comparative survey of the wearable and portable assistive devices for visually-impaired people in order to show the progress in assistive technology for this group of people”. Katzschmann *et al.,* (2018) developed a “Safe Local Navigation for Visually Impaired Users with a Time-of-Flight and Haptic Feedback Device”. The proposed system presents “ALVU (Array of Lidars and Vibrotactile Units), a contactless, intuitive, hands-free, and discreet wearable device that allows visually impaired users to detect low- and high-hanging obstacles, as well as physical boundaries in their immediate environment”.

Considering the development of a navigation device capable of guiding the blind through indoor and outdoor scenarios a major challenged; (Santiago Real & Araujo., 2019) went into a research on “Navigation Systems for the Blind and Visually Impaired: PastWork, Challenges, and Open Problems”. Their objective is

to “provide an updated, holistic view of this research, in order to enable developers to exploit the different aspects of its multidisciplinary nature”. However, smartphones and similar tools are rapidly making their way into the daily routines of Blind and Visually Impaired. In this context, old and novel solutions have become feasible, some of which are currently available in the market as smartphone applications or portable devices.

Having seen how crucial it is for the visually impaired persons to live independently, Vijeesh *et al.,* (2019) carried out a research on “Design of Wearable shoe for blind using LIFI Technology”. they researchers “proposed a system that hypothesizes a smart shoe that alerts visually impaired people over obstacles that could help them to walk independently with low chances of accident”. To overcome the challenges associated with the smartcane, “a smart shoe is designed and developed to detect obstacles in the path of visually impaired person which will make them feel like a normal individual. Garcia-macias *et al*., (2019); Bai *et al.,* (2019) also developed a wearable device to enhance mobility of the visually impaired,

Islam *et al*., (2019) went into a research on “Developing Walking Assistants for Visually Impaired People: A Review”. The researchers state that “due to the rapid growth of People with visual impairment in recent decades, the development of walking assistants for visually impaired people has become a prominent research area. This review discusses the recent innovative technologies developed for the visually impaired to aid them in walking with their merits and demerits. With the help of this review, a schema is drawn for upcoming development in the field of sensors, computer vision, and smartphone-based walking assistants. The goals of the developed system are to generate an audio signal and/or vibration with the presence of obstacles for indoor and outdoor surroundings. Evidently, the actual

goal is achieved when physicians and computer scientists develop an assistive technology for visually impaired people in the future”

2.9 **Related works on Text Detection and Recognition**

Babu *et al.*, (2015); Kevadia *et al.,* (2018), developed a method for recognizing printed text, detecting Quick Response Code (QR) and recognizing faces in arbitrarily acquired images. However, the system is basically for optical character recognition, text to speech conversion and face recognition. The concept of voice recognition was not implemented.

Lin *et al.*, (2017) tried to solve navigation problems for the visually impaired via smartphone Integrated computer image recognition system and smartphone application to form a guiding system”.

Tejaswani *et al.,* (2018) argued that the braille system of learning and braille books specifically for the visually impaired are not available everywhere especially in the rural areas, as a result of this, it has become very difficult for the visually impaired to access text and print documents. To overcome this difficulties faced by the visually impaired, “finger reader with speech assistance is used to read any printed text from the mobiles, newspapers etc”. The Finger Reader warns the readers when they move away from the line. However, Finger reader only reads the text and does not specify whether it is headline or sub head line. Also, the camera does not auto-focus it takes time to focus on the text.

Muthusenthil *et al.,* (2018) worked on “Smart Assistance for Blind People using Raspberry Pi”. This research “addresses the integration of a complete Text Read-out system designed for the visually challenged. The system consists of a webcam interfaced with raspberry pi which accepts a page of printed text. However, the

system is purely a text to speech conversion and may not function well for other purposes like voice and face recognition.

Having seen communication as something that can be done through speech or text, Kiran and Chitera (2019) proposed a novel implementation of smart guiding system for the visually challenged. The system is said to be a novel and real-time cost beneficial technique. It allows the visually impaired to hear the contents of the text images instead of reading through them.

Kim & Park (2019) affirms that there are existing assistive devices and aids for visually impaired people but these devices have some functional limitations like not being able to process multimedia contents such as images and videos. Based on these functional limitations, they proposed a system that "describes a methodology to effectively convert multimedia contents to braille using 2D braille display, transforms Digital Accessible Information System (DAISY) and electronic publication (EPUB) formats into 2D braille display and introduces interesting research considering efficient communication for the VI.

Persons". (Kumar & Munindhar, 2019) developed an assistive system that will enable the visually impaired gain increased independence and freedom in the society.

Pruthvi *et al.*, (2019) designed a system to help guide the visually impaired by detecting objects and portray the information to them in the form of speech.. (Elgendy *et al.,* 2019) in their research work titled "Making Shopping Easy for People with Visual Impairment Using Mobile Assistive Technologies", provides an overview of the various technologies that have been developed in recent years to assist people with visual impairment in shopping tasks. Their study also introduced latest direction in this area, which will help developers to incorporate such

solutions into their research. “This study has discussed the current, most prevalent, solutions to help PVI shop effectively.

(Chen *et al.,* 2019) proposes a smart wearable system that performs image recognition in their research titled “A Novel Approach to Wearable Image Recognition Systems to Aid Visually Impaired People”. The system adopts the method of cloud and local cooperative processing.

(Mollah *et al.*, 2011), designed “an Optical Character Recognition System for Camera-based Handheld Devices”. At first, they extracted text regions and perform skew correction. After which, the extracted regions were binarized and segmented into lines and characters. They achieved a maximum recognition accuracy of 92.74%. from the experiment conducted with 100 set of business card images. With this, they affirmed that the result of their technique is computationally more efficient than using Tessseract and it consumes low memory, which made it suitable for handheld devices.

(Wang *et al*., 2013) conducted a study on “Natural Scene Text Detection with Multi-channel Connected Component Segmentation”. The research here confirmed that “though text detection is a research that with a very high interest and attention till date, yet, natural scene text detection still poses great challenges and problem due to the text variations and the complexity of the background”. As a way of confronting these challenges, they designed an “efficient text detection method with multi-channel connected component segmentation. This was achieved by the use of Markov Random Field with local contrasts, colors and gradients of RGB channels. They evaluated their system with ICDAR 2003 dataset and the ICDAR 2011 dataset it was proven to have favorable comparison with the state-of-the-art methods”. However, their method fails in some natural scene text images with low contrast text, strong illumination and non-horizontal text.

(Uma & Dagade, 2018) carried out a study on "Maximally Stable Extremal Region Approach for Accurate Text Detection in Natural Scene Images", with the objective of "designing a text detection system which will be able to detect maximum characters from natural scene images. They extracted character candidate from the binarized text using MSER region detector algorithm". From their result, the MSER method proved better results from other methods of text detection and recognition. However, detecting highly blurred texts in low-resolution natural scene images was a difficult one for their approach.

(Karande & Jogdand, 2018) developed a "Portable Camera-Base Assistive Text Reading from Hand Held Objects for Blind Person". Their study was "aimed at enabling the blind to access text labels from hand-held objects as they go about their daily activities. This was achieved by first capturing the labels with camera, processed the captured images using text localization and recognition algorithm and extracted text using MATLAB". Finally, the extracted text label was made available to the blind person in audio format. However, their work was limited to label reading and hence, should be enhanced to detect text in different structures, scene images and documents with complex background and multipart.

(Trémeau *et al.*, 2011) proposed "a novel method for detecting and segmenting text layers in complex images. The proposed system is a robust system that is capable of handling degradations like shadows, uneven illumination, low-contrast, large signal dependent noise, smear and strain. This was achieved by the use of Generalized Extreme Value Distribution (GEVD) to find a correct threshold for binarization".

Still on the quest to confront the challenges associated with scene text detection and recognition, (Yi & Tian, 2012) developed and an "Assistive Text Reading

from Complex Background for Blind Persons". This is to enable them read text from signage and objects that are held in the hand. Their system was developed following Adaboost model. Off-the- shelf optical character recognition (OCR) software was used to recognize Text characters in the localized regions and transformed into speech outputs. The system was evaluated on ICDAR 2003 Robust Reading Dataset.

Shah, & Jain (2015) proposed a camera based prototype system for extracting from the image the label of the product which a blind user wishes to buy from and will be pronounced as audio output to the blind user. Their system uses background subtraction to select the object of interest and the system was implemented on Raspberry Pi board.

Baek *et al*., (2019) having critically considered the challenges associated with traditional scene text detection techniques, propose a novel text detector using CRAFT, that localizes the individual character regions and links the detected characters to a text instance.

Kambala et al., (2020) developed an OCR system for a web application. To achieve this, they "implemented a hybrid model containing three components namely, the Convolutional Neural Network component, the Recurrent Neural Network component and the Transcription component which decodes the output from RNN into the corresponding label sequence". The web application developed simulates an optical character recognition system by providing the functionality of text recognition. Their study was indeed a novel one which can be enhanced by outspreading it with certain improvements such as applying state-of-the-art text detection techniques and also making it real time application.

Shikha *et al*., (2020) developed an Ancient Text Character Recognition Using Deep Learning. Their study basically focused on a model for ancient character text recognition and translating it into text that could be used further for translating into recognizable language. Contribution to knowledge was to perform image recognition of the low resource Sundanese language and convert it into text language and further convert it into universally accepted language. However, the proposed approach was limited to one language (Sundanese) and could be conducted on multilingual ancient scripts recognition system In future.

Hanaa Fathi Mahmood, (2020) also carried out a study on development of Text detection and Recognition from Natural Images. "They proposed a methodology to handle the problem of text detection by using novel combination and selection features to deal with the classification algorithms of the text/non-text regions. They extracted the text-region candidates from the grey-scale images using the MSER technique. While the initial detection was refined and validated using machine Method. The proposed methods as affirmed by the researchers achieved good detection and recognition results in different. However, their framework can be extended for application to videos and GIF files".

2.10 **Closely Related Works**

Recently, many researchers had made some progress on analyzing and improving the challenges confronting the VIP in the area of obtaining and maintaining employment which centered on mobility and navigation, having access to print documents, multipart and complex background document and scene images. To enable building an enhanced text reading system for the visually impaired individuals, the following existing systems were thoroughly analyzed:

1) **Smart Assistance for Blind People using Raspberry Pi by (Muthusenthil, *et al.* 2018)**

Muthusenthil, *et al. (*2018) designed a system that will assist the visually impaired to identify and differentiate objects of the same kind through accessing printed Document tagged on the objects. From their findings, the number of tags involved in the design can easily be expanded since the system only does scanning and reading aloud the content of a tag, there is no database involved in storing names of objects. The system "consists of a webcam interfaced with raspberry pi which accepts a page of printed text".

The designed system is the "integration of a complete Text Read-out system designed for the visually challenged. The system consists of a webcam interfaced with raspberry pi which accepts a page of printed text. The OCR (Optical Character Recognition) package installed in raspberry pi scans it into a digital document. After which, the name of the detected object is read aloud to the visually impaired by a text to speech conversion unit (TTS engine) installed in raspberry pi". The output is fed to an audio amplifier before it is read out as shown in figures 2.20 and 2.21.

Limitations of Muthusenthil, *et al. (*2018)

The system proposed by Muthusenthil, *et al. (*2018) has the following limitations:

1. the system is purely an Image to text and then speech conversion
2. Identification and decision is based on the tags placed on each of the items and not on available data
3. the system lacks Smartness and intelligence
4. The system is surrounded with lots of discomforts and inconveniences since the Raspberry Pi (a hand held device) has to be moved very close to the item for identified before capturing is done.

5. Communicate must be through loud speaker. There was no room for wireless means such as Bluetooth and ZigBee technology.

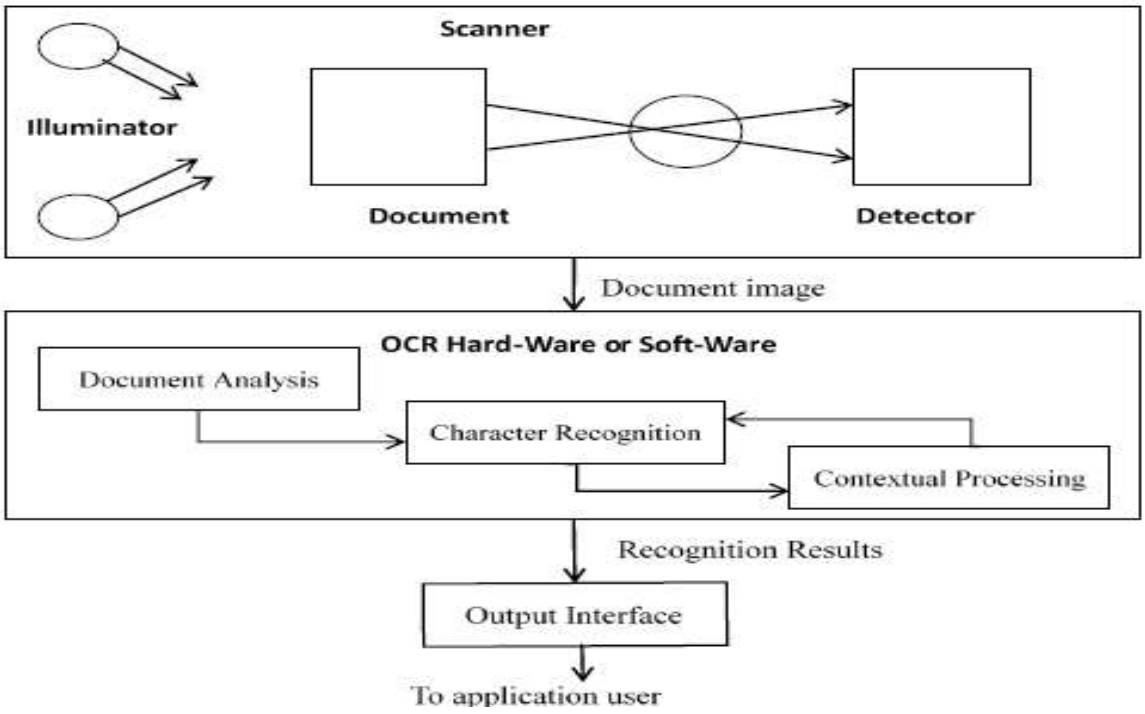


Figure 2.20: Architecture of Muthusenthil, *et al. (*2018)

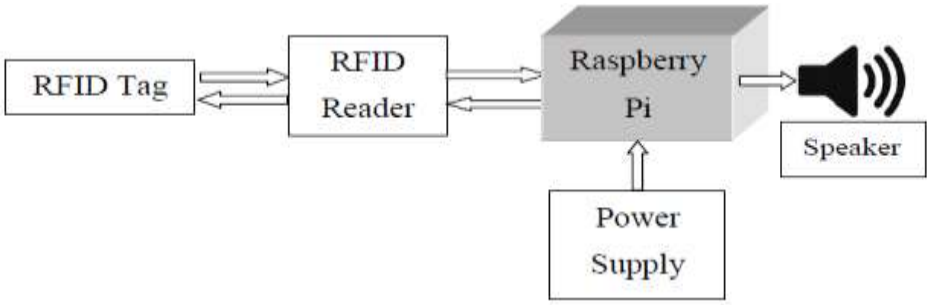


Figure 2.21: Block Diagram of Muthusenthil, *et al. (*2018)

2). Portable Camera Based Assistive Text Reading from Hand Held Objects for Blind Person by (Karande & Jogdand, 2018)

(Karande & Jogdand, 2018) developed a system that will enable the blind people to access labels on handheld devices. This was achieved by first capturing the labels

with camera, processed the captured images using text localization and recognition algorithm and extracted text using MATLAB. Finally, the extracted text label was made available to the blind person in audio format.

Limitation of (Karande & Jogdand, 2018)

Their work was restricted to label reading as shown in figure 2.22 and hence, should be enhanced to detect text in different structures, scene images and documents with complex background and multipart. Samples of handheld objects and the flowcharts of the system are shown in figure 2.22 and 2.23 accordingly.

Figure 2.22: Examples of printed text from hand-held objects. *Source* (P. Karande, S. Jogdand, 2018)

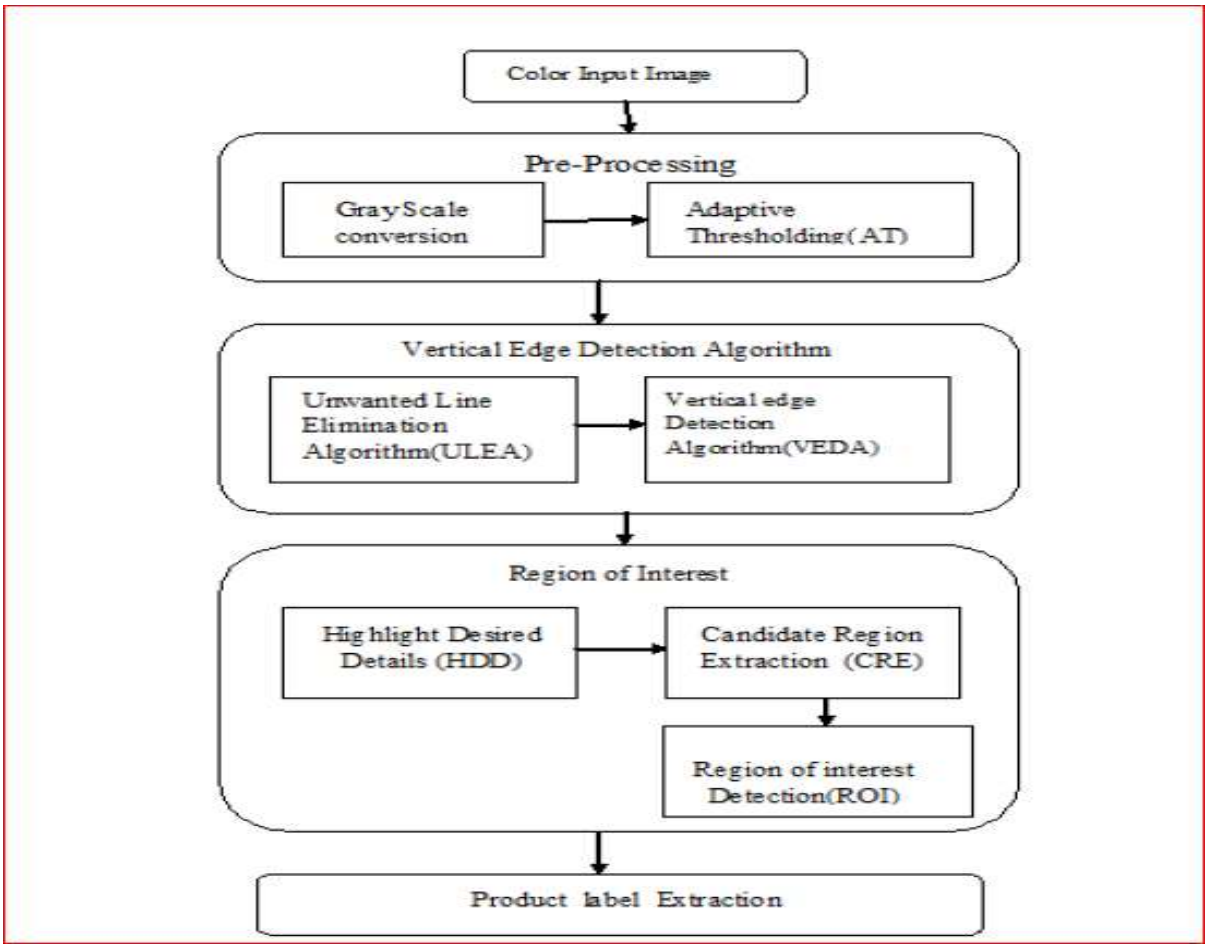


Figure 2.23: Flowchart of (Karande & Jogdand, 2018)

3). Analysis of Maximally Stable External Region Approach for Accurate Text Detection in Natural Scene Images by (Uma & Dagade, 2018)

Uma & Dagade, (2018) carried out a study on the "Analysis of Maximally Stable External Region Approach for Accurate Text Detection in Natural Scene Images". In their system, Otsu binarization algorithm was applied on the input color image so as to be converted to gray or black and white background. After which Maximal Stable Extremal Region (MSER) Detector algorithm was used to extract character candidate from the binarized image. The architecture of (Uma & Dagade, 2018) is shown in figure 2.24.

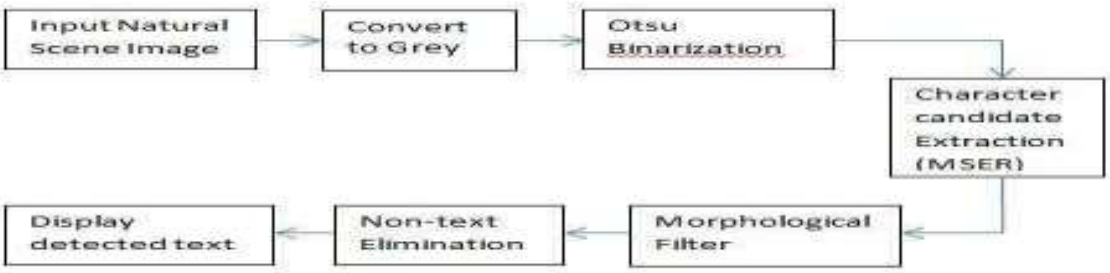


Figure 2.24: Architecture of (Uma & Dagade, 2018)

Limitation of (Uma & Dagade, 2018)

Their system uses MSER and Otsu Binarization algorithm which has the following limitations:

i. Their system was unable to "detect highly blurred texts in low-resolution natural scene images
ii. It does not perform well for images with uneven illumination and shadow
iii. It gives its best performance for only those images that have clear bi-modal pattern". Unfortunately, degraded documents normally don't have such clear cut pattern. Hence, the performance of the system is affected when confronted with uneven illumination, poorly scanned documents and shadow.

4). Ancient Text Character Recognition Using Deep Learning (Shikha *et al.*, 2020)

Shikha et al., (2020) "proposed a character level recognition approach for Sundanese text written on palm leaf using deep learning approach. The available script is in degraded format which needs to be preserved and converted into text format for future generations. For this purpose, they proposed an encoder decoder technique for binarization and their results shows binarization accuracy of 74.24% and F-measure to be 75%. However, the recognized images use one to one

mapping for further translation into recognizable text i.e. English". Future work should be conducted on multilingual ancient scripts recognition system, as the approach proposed is limited to one language i.e. Sundanese. Figure2.25 shows the Shikha *et a*l., (2020). From the Architecture, characters are recognized using a three layer Convolutional Neural Network.

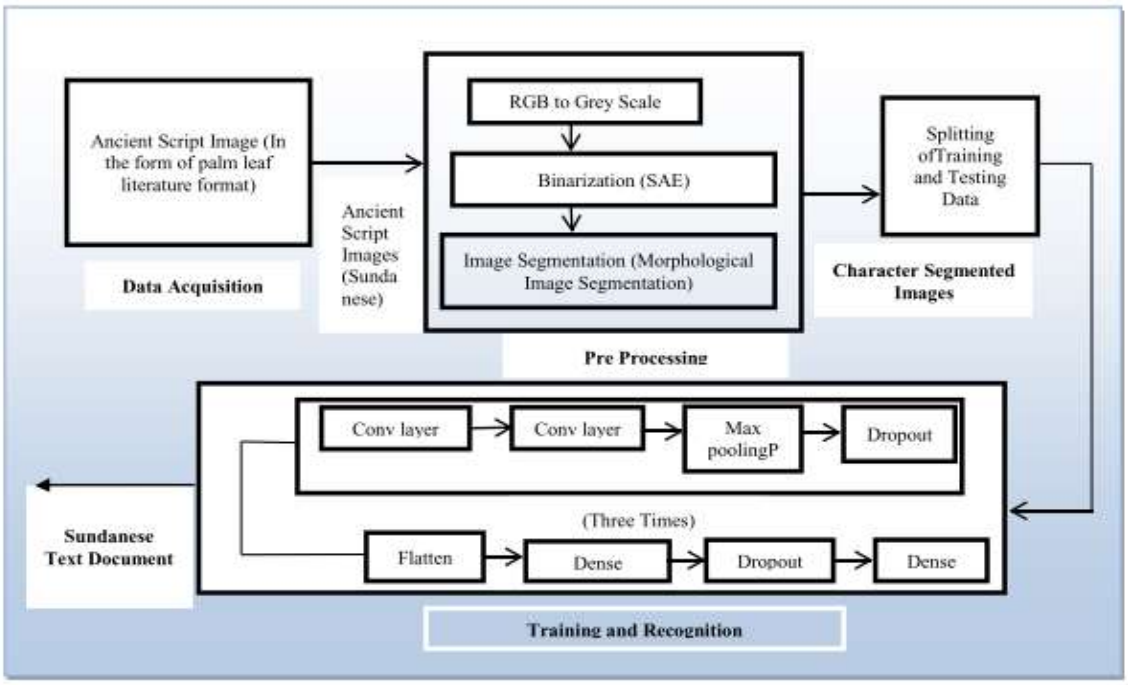


Figure 2.25: Architecture of Shikha et al., (2020).

2.11 Gap Statement

Research has shown that processing text elements in documents with non-regular or complex mix of elements and quality issues still poses computational difficulties to OCR systems because texts can be written in different fonts, appear in different resolutions, appear with irregular spacing, overlap and bear artistic effects (Fei *et al.*, 2018 ; Kambala *et al.*, 2020 ;Shikha et al., 2020). Further, camera-based OCR are susceptible to lighting conditions and other operator dependent exposure settings that often result in blurred or perspective-distorted images and other forms

of degradation such as warping and shadow overlay (Mhaske *et al* 2016; Trémeau *et al*., 2011; Hanaa Fathi Mahmood, 2020).

High dimensionality of the feature space is one of the very prominent challenges for OCR techniques on the problem scenario described in the preceding paragraphs. The problem of high dimensionality is mostly occasioned by the presence of noise in the form of a subsuming background or low contrast background which reduces the accuracy of text detection and recognition by OCR systems. Thus the computational time is increased without corresponding improvement in recognition rate.

Therefore, having critically reviewed some of the existing works like (Uma, & Dagade, 2018; Karande & Jogdand, 2018; Hanaa Fathi Mahmood, 2020; Shikha et al., 2020), it was noted that Traditional machine learning approaches to text detection and recognition face challenges when confronted with:

iv. highly blurred texts in low-resolution natural scene images
v. Documents with uneven illumination and shadow
vi. Degraded documents and Curved images

They actually give their best performance for "only those images that have clear bi-modal pattern". Unfortunately, degraded documents normally don't have such clear cut pattern.

2.12 Direction of Research

The literatures reviewed in this chapter gave a comprehensive view of countless smart and assistive devices for the visually impaired using advanced assistive technologies. Nevertheless, having critically reviewed a handful of closely related works, it was observed that many researchers had made frantic efforts to ensure

independent living for people with visual impairment especially in the area of obstacle detection and avoidance, indoor and outdoor navigation system, barriers in obtaining and maintaining employment, shopping assistance and accessing Print documents, none has been able to totally eradicate the challenges of Visually Impaired persons.

Though several systems had been designed to improve the quality of life of people with visual incapacitations and also ensure their independent mobility. The traditional flat-bed OCR is also noted to be fast with some level of accuracy, most of them are not capable of recognizing text in scene mages. Contrary to making life easy and economical, some of these systems already in existence are very expensive thereby making it inaccessible for the target group.

Hence, this research on Robust Camera-Based Text Recognition model for the visually impaired is geared toward providing affordable and accessible text recognition system that will grant the VIPs access to print documents and scene images.

To align with driving the culture of independent living for the blind, this research therefore proposes to explore the first order edge detection , precisely the canny edge detection techniques and the deep learning approaches for the improvement of character recognition accuracy in regular, multipart document and in scene images using a hand-held camera. The chosen approaches were based on the following reasons, they are:

1. Are adaptive in nature
2. Tolerates Noise more than the classical operators.

3. Removes streaking Problems by applying hysteresis techniques which uses two threshold values t_{low} and t_{high} . Unlike the classical operators that uses single threshold and thus finds it difficult to handle streaking.
4. Enhanced Localization

This system if successfully implemented and adopted, will give the visually impaired the opportunity of being gainfully employed and working comfortably in an office

CHAPTER THREE

MATERALS AND METHODS

3.1 Preamble

Having revealed in the problem statement and preceding chapters the challenges associated with detecting and recognizing scene text, it is paramount that a model be developed for enhanced text recognition system. Since data is the bed rock of deep learning and the efficacy of every learning algorithm is determine by the quality of the training data set used, this chapter therefore discussed the materials used and the steps taken to achieve the proposed system.

The model consists of the following modules:

i. Data acquisition module that consists of primary data acquired using attached camera and secondary data collection from existing data set.
ii. Text localization and detection module. This involved Pre-processing to remove noise and complex background from the mage and converting into grey scale.
iii. Text recognition module that recognizes the text and converts it into text format.
iv. The text to speech module which translate the recognized text into speech and finally communicate to the VIP in audio format.

3.2 **Primary Data Acquisition**

In primary data acquisition module image is acquired using attached camera. The captured mage is sent to the system's inbuilt scanner which detects and describes

local features in the image. The detected features are then analyzed for further preprocessing. Outlined next are the steps involved in image acquisition.

1. ***Set webCamFeed to True, and create a cap variable to capture imaging using external camera (0). Also set image dimension to 640 x 480 pixels***
2. ***Initialize trackbars for adjusting upper and lower threshold for Canny Edge detector***
   (a) create function to initialize trackbar
3. ***Set camera in an infinite loop to make for continuous reading of frames***
4. ***Resize image, convert to grayscale and apply Gaussian Blur***
5. ***Get trackbar thresholds***
   (a) create function to validate trackbars
6. ***Apply Dilation and Erosion to image***
7. ***Find and draw all detected contours***
8. ***Find the biggest contour:***
   (a) create function to find biggest contour in contours
   (b) create function to reorder biggest contour
   (c) create function to draw rectangle around biggest contour
9. ***Prepare points for warp perspective to straighten image***
10. ***Remove 20 pixels from each side of image***
11. ***Apply Adaptive Thresholding***
12. ***Create array for various image transformations for display***
13. ***Create labels for displayed images***
14. ***Stack images***
    (a) create function to stack images
    (b) Save image to file Scanned

3.3 Preprocessing and Text Localization

The second phase after capturing is preprocessing: the main objective of this phase is to make it easy and possible for the OCR to distinguish character/word from the background. This was achieved using the CRAFT Algorithm which incorporates

the following Binarization, skew correction, Noise removal and Thinning and Skeletonization techniques. This is done to ease the job of text localization and detection. Outlined next is the Pipeline for Developing Text Detection.

3.3.1 Pipeline for Developing Text Detection Model

To develop text detection model, CRAFT Technique was adopted. This technique outperforms the traditional approaches of text detection like MSER which finds it difficult to handle curved text, poorly scanned text, very long and degraded or deformed text. Notwithstanding, MSER algorithm is sensitive to blurred image so gradient amplitude is added to it to produce edge-preserving MSER (Abraham, 2015 ; Uma & Dagade, 2018; Mahmood, 2020). Therefore, to effectively detect text areas by exploring each character and affinity between characters, CRAFT which adopts a fully convolutional neural network was designed. Its final output produces two channels as Score Maps: the Region Score which localizes individual character in the image and the Affinity Score which groups each character into a single instance. Afterwards, the affinity and region scores are combined to give the bounding box for each word in the image (Baek *et al*., 2019). Figure 3.3 shows an illustration of CRAFT architecture. The pipeline for this architecture is as follows:

1. Transformation of the Captured Scene Image

This involves all the pre-processing done on the image after capturing to prepare it for the text localization/detection and recognition stages. The algorithm for Text detection model following CRAFT approach is as shown below:

a. **CRAFT algorithm for Text Transformation**

(a) Create a function to load image from directory, resize and return grayscale image:
(b) ***fetch data from directory***
(c) ***resize image***
(d) ***convert image to grayscale***

(e) ***apply Gaussian Blur***
(f) ***binarize Image using OTSU Thresholding and BINARY INVERSION***
(g) `return image`

b. Text LOCALIZATION/DETECTION

Text Detection as earlier stated in Chapter two is the process of predicting and localizing text instances from the image. Having carried out extensive literature review and critically analyzed some of the closely related literature such as **(**Muthusenthil, *et al.*2018; Karande & Jogdand, 2018; Uma & Dagade, 2018; Baek et al., 2019 ; Mahmood, 2020), it was concluded that Scene Text Detection Based on Deep Learning Approaches precisely character region awareness for text detection (CRAFT) algorithm is best suited for all kinds of text regardless of the text length and shape. It adopts fully convolutional network architecture with a VGG-16 backbone Figure 3.1 shows an illustration of CRAFT architecture.

The CRAFT architecture as discussed in chapter two predicts two scores for each character in image string:

Region Score: localizes individual character in the image.

Affinity Score: this is used to merge/group characters into a single instance (a word).

the affinity and region scores are combined to give the bounding box for each word in the image (Baek et al., 2019) as shown in Figure 3.1 (Baek *et al.*, 2019)

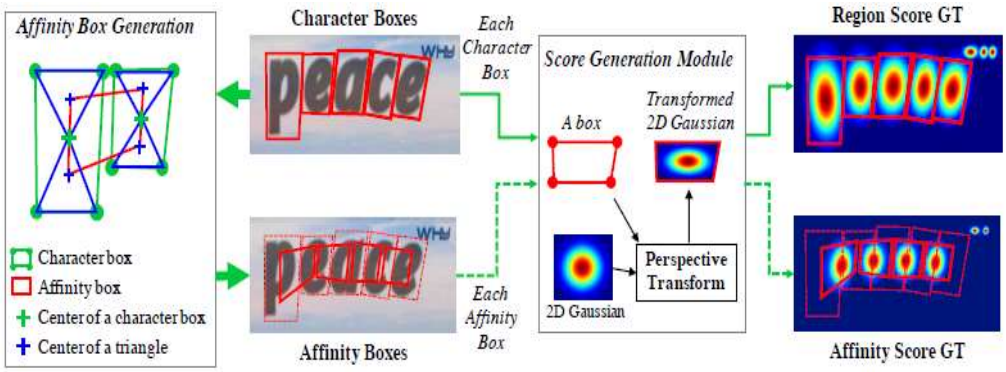


Figure 3.1: Illustration of the General Training stream of CRAFT

c. CRAFT algorithm for Text LOCALIZATION

a. create functions to get boxes:

```
b. initialize the craft model, and load weights from pretrained model
c. save score text to file
```

d. CRAFT algorithm for Text DETECTION

a. import modules, weights from pretrained model and vgg16 backbone

```
b. create class for double convolutional neural network
c. create the CRAFT class
d. create the CRAFT model using the CRAFT class. Set pretrained to TRUE to
   imported weights from the craft_mlt_25k.pth pretrained model
```

e. CRAFT Model For Text Localization And Detection

To generate the ground truth Orientation Map
the upward angle of the GT character bounding box is represented as θ^{*box}

the channel predicting x, y-axis has a value of

$$S^{*}cos(p) = (\cos\theta + 1) \times 0.5 \quad S^{*}sin(p) = (\sin\theta + 1) \times 0.5 \qquad (15)$$

The loss function for orientation map is calculated by (16)

$$L_\theta = S^{*}{}_r{}^{(p)} \cdot (||Ssin(p) - S_{*}sin(p)||_2^2 + ||Scos(p) - S_{*}cos(p)||_2^2)$$

*Where S*sin(p) and S*cos(p) denote the ground truth of text orientation*

The character region score $S_r(p)$ is used as a weighting factor

Objective function in the detection stage L_{det} *is defined as,*

$$L_{det} = L_r + L_l + \lambda L_\theta \qquad (17)$$

Where Lr and Ll denote character region loss and link loss

the angle of the text box is calculated by:

$$\theta_{box} = \arctan\left(\frac{\sum(S_r(p)\times(S_{\sin}(p)-0.5))}{\sum(S_r(p)\times(S_{\cos}(p)-0.5))}\right) \qquad (18)$$

3.4. Proposed System's Architecture

The architecture of the proposed system is given in figure 3.2. It described in details all the different stages and steps deployed to achieving a Camera-Based text Recognition model for the visually impaired. It shows the process and approaches

deployed to achieving image acquisition, Pre-processing and image transformation, text localization and detection, feature extraction and text recognition. Finally the, the detected, identified and recognized text is communicated to the blind user after text to speech transformation via audio format.

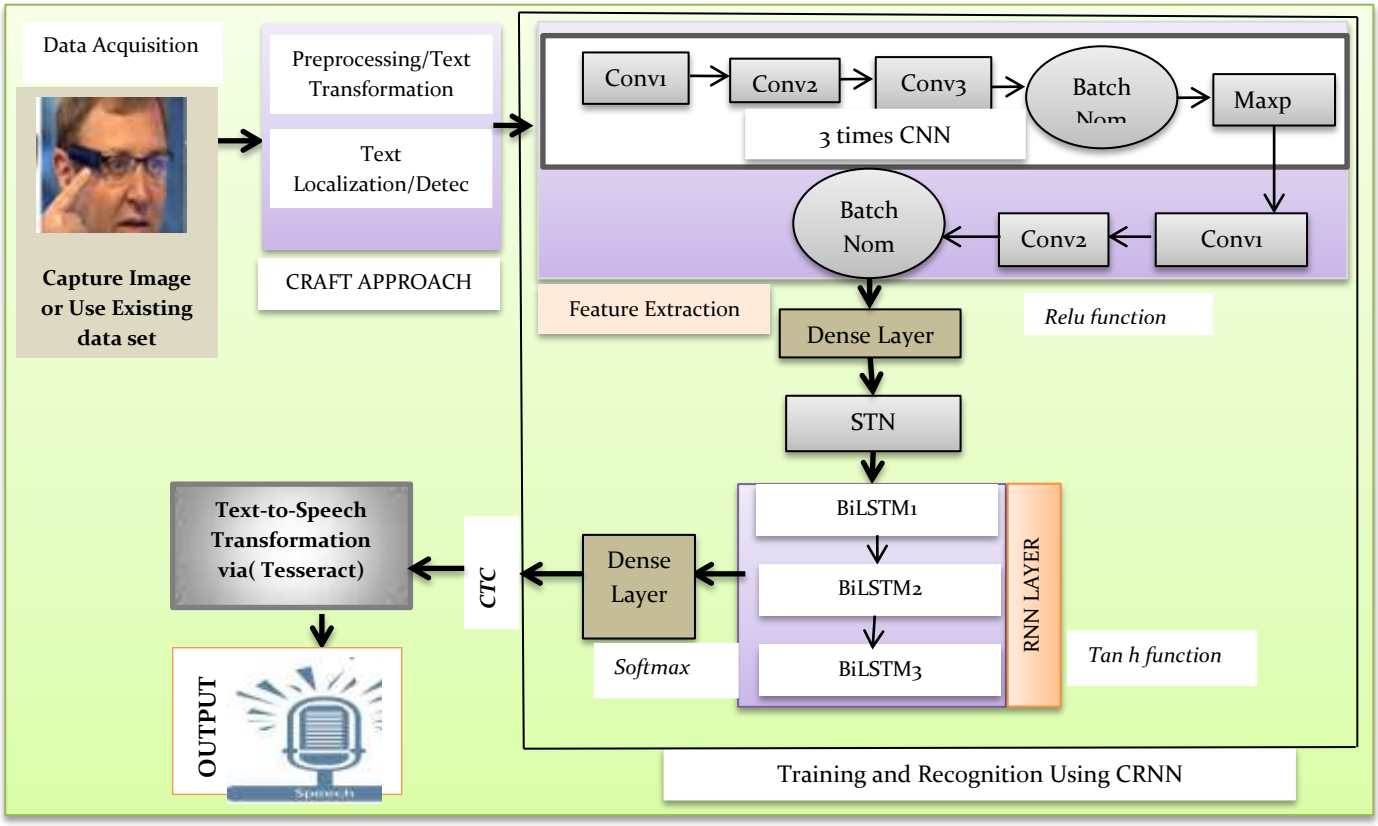


Figure 3.2**:** Architecture of the Proposed System

3.5. Text Recognition/Extraction Model

Having developed a model for achieving text localization and detection as stated in objective one above, the next phase is the recognition and extraction part. To achieve this, Convolutional Recurrent Neural Network (CRNN) and Connectionist Temporal Classification (CTC) Loss was adopted. CRNN is a combination of CNN, RNN, and CTC (Connectionist Temporal Classification) loss for image-based sequence recognition tasks, such as scene text recognition and OCR. The

architecture of the convolutional layers is based on the VGG-VeryDeep architectures. The architecture of CRNN consists of convolutional layers which extract feature maps automatically and a recurrent network used for predicting the text from each frame of the feature map sequence. As noted in chapter two also, the implementation of CRNN follows 6 steps (collecting data, preprocessing, network architecture, loss function, model training and model testing). These steps as outlined next were followed to build a more efficient recognition model.

3.5.1. Collecting Dataset

Model training was done using the ***Synth90k*** synthetic text dataset provided by the Visual Geometry Group. It has 9million synthetic images generated with a dictionary of 90k English words.

3.5.2. Preprocessing the Existing Data Set (Primary Data).

Now, having obtained dataset the next step is to make it acceptable for the proposed model by preprocessing both the input and the output labels. The steps outlined next are followed:

a) Read the image and convert into a gray-scale image
b) Make each image of size (128,32) by using padding
c) Expand image dimension as (128,32,1) to make it compatible with the input shape of architecture
d) Normalize the image pixel values by dividing it with 255.

For the output labels the steps below are taken.

a) Read the text from the name of the image as the image name contains text written inside the image.

b) Encode each character of a word into some numerical value by creating a function (as ‘a’:0, ‘b’:1 ‘c’:2,’d’:3,…….. ‘z’:26 etc). Let say we are having the word ‘cdcd’ then our encoded label would be [2,3,2,3]

c) Compute the maximum length from words and pad every output label to make it of the same size as the maximum length. This is done to make it compatible with the output shape of the RNN architecture.

The outcome of the preprocessing stage is as outlined next.

a. Data Preparation Steps

#1. Fetch the whole synth90k dataset and unzip 15GB of the total 21GB:

```
!curl -LO https://thor.robots.ox.ac.uk/~vgg/data/text/synth90k.tar.gz
!tar -xf synth90k.tar.gz
```

#2. Create three variables to hold for training, validation and test datasets and set header to None:

```
train = pd.read_csv(dir_path+'/annotation_train.txt', header = None)
validation = pd.read_csv(dir_path+'/annotation_val.txt', header = None)
test = pd.read_csv(dir_path+'/annotation_test.txt', header = None)
```

#3. Check for the length of training, validation and test variables:

Original Train Dataset Size : 7224612

Original Validation Dataset Size : 802734

Original Test Dataset Size : 891927

#4. Output format of dataframe in train variable:

Table 3.1: Output format of dataframe in train variable

	0
0	./2425/1/115_Lube_45484.jpg 45484
1	./2425/1/114_Spencerian_73323.jpg 73323
2	./2425/1/113_accommodatingly_613.jpg 613
3	./2425/1/112_CARPENTER_11682.jpg 11682
4	./2425/1/111_REGURGITATING_64100.jpg 64100

#5. Create actual directories for the Dataset and to hold train, validation and test datasets:

```
# Creating the Directories
! mkdir Data
! mkdir Data/train
! mkdir Data/validation
! mkdir Data/test
```

#6. Move image paths into their designated folders in the Data folder:

```
for path in tqdm(train['path'].values):
    if os.path.isfile(path)==False:
      pass
    else:
      os.rename(path, 'Data/train/'+path.split('/')[-1])

for path in tqdm(validation['path'].values):
  if os.path.isfile(path)==False:
      pass
  else:
      os.rename(path, 'Data/validation/'+path.split('/')[-1])

for path in tqdm(test['path'].values):
  if os.path.isfile(path)==False:
      pass
  else:
    os.rename(path, 'Data/test/'+path.split('/')[-1])
```

#7. Extract labels embedded in path and create a dataframe for both paths and labels:

Table 3.2: Extracted Labels Embedded in Parts

	Path	Label
0	Data/train/57_babes_5257.jpg	BABES
1	Data/train/358_twinkies_81407.jpg	TWINKIES
2	Data/train/222_BIOCHEMICALS_7556.jpg	BIOCHEMICALS
3	Data/train/387_flecked_29483.jpg	FLECKED
4	Data/train/403_CHATLINES_12870.jpg	CHATLINES

b. Data Cleaning Steps

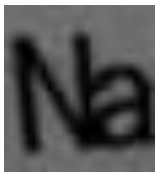
#1. We check the dataset for NA values and confirm whether they are string errors or match image in the image path.

Table 3.3: Checking the data based for NA Values

	Path	Label
36954	Data/train/74_NULL_52538.jpg	NaN
101987	Data/train/215_NULL_52538.jpg	NaN
120953	Data/train/64_NULL_52538.jpg	NaN
174857	Data/train/249_NULL_52538.jpg	NaN
179073	Data/train/277_Null_52538.jpg	NaN
200061	Data/train/7_NULL_52538.jpg	NaN
219402	Data/train/434_Na_50735.jpg	NaN
314357	Data/train/168_Null_52538.jpg	NaN
315107	Data/train/93_NULL_52538.jpg	NaN
329920	Data/train/194_na_50735.jpg	NaN

Checking one of the NA labels for matching image:

```
img = Image.open(train.path[656460])
img
```



We find that the NA values point to matching images in the path. Hence, we set NA filter to False in the data-fetching procedure:

```
train = pd.read_csv('Data/train.csv', na_filter=False)
validation = pd.read_csv('Data/validation.csv', na_filter=False)
test = pd.read_csv('Data/test.csv', na_filter=False
```

#2. We analyze label length for train, validation and test images using Count Plot and Cumulative Distribution Function (CDF) as shown in figure 3.3 and figure 3.4.

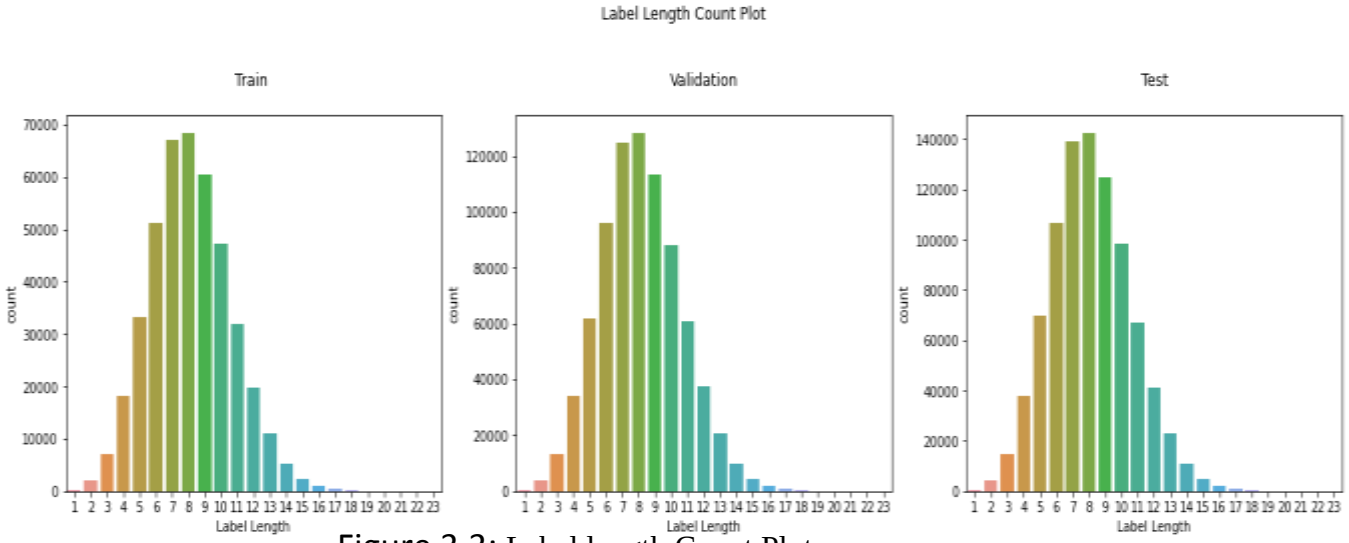


Figure 3.3: Label length Count Plot

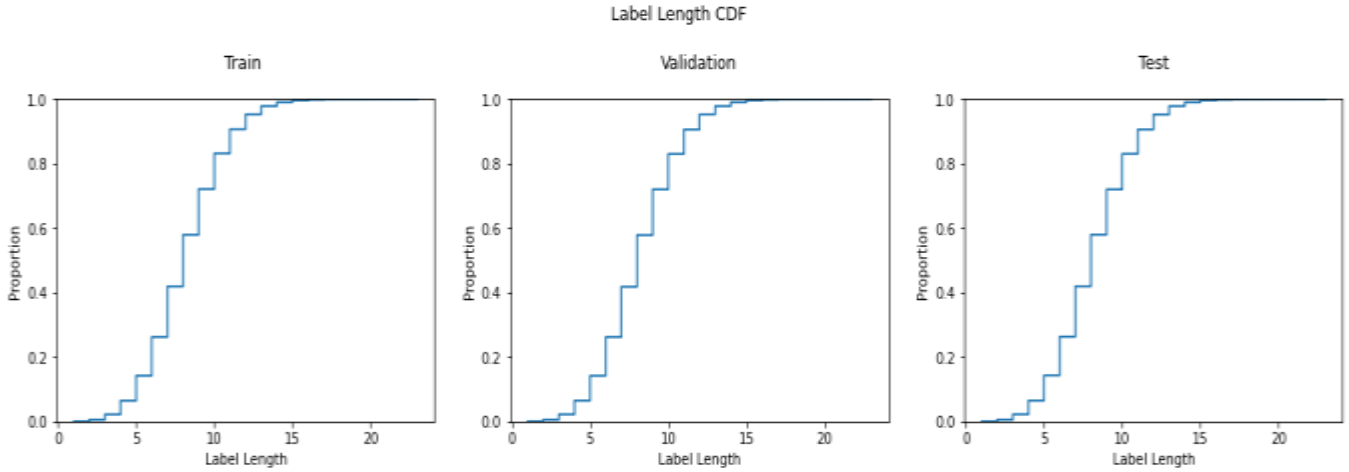


Figure 3.4: Label CDF

3. We check for the number of Numerical and String data in our Label:

```
train_num_labels = []
train_non_num_labels = []

for idx,l in enumerate(tqdm(train.label)):
    if(l.isnumeric()):
        train_num_labels.append(idx)
    else:
        train_non_num_labels.append(idx)

validation_num_labels = []
validation_non_num_labels = []
```

```
for idx,l in enumerate(tqdm(validation.label)):
    if(l.isnumeric()):
        validation_num_labels.append(idx)
    else:
        validation_non_num_labels.append(idx)

test_num_labels = []
test_non_num_labels = []

for idx,l in enumerate(tqdm(test.label)):
    if(l.isnumeric()):
        test_num_labels.append(idx)
    else:
        test_non_num_labels.append(idx)
```

#4. Analyze image height and widths for the train, validation and test dataset using Count Plot and Cumulative Distribution Function. This is as shown in figures 3.5, 3.6, 3.7 and 3.8.

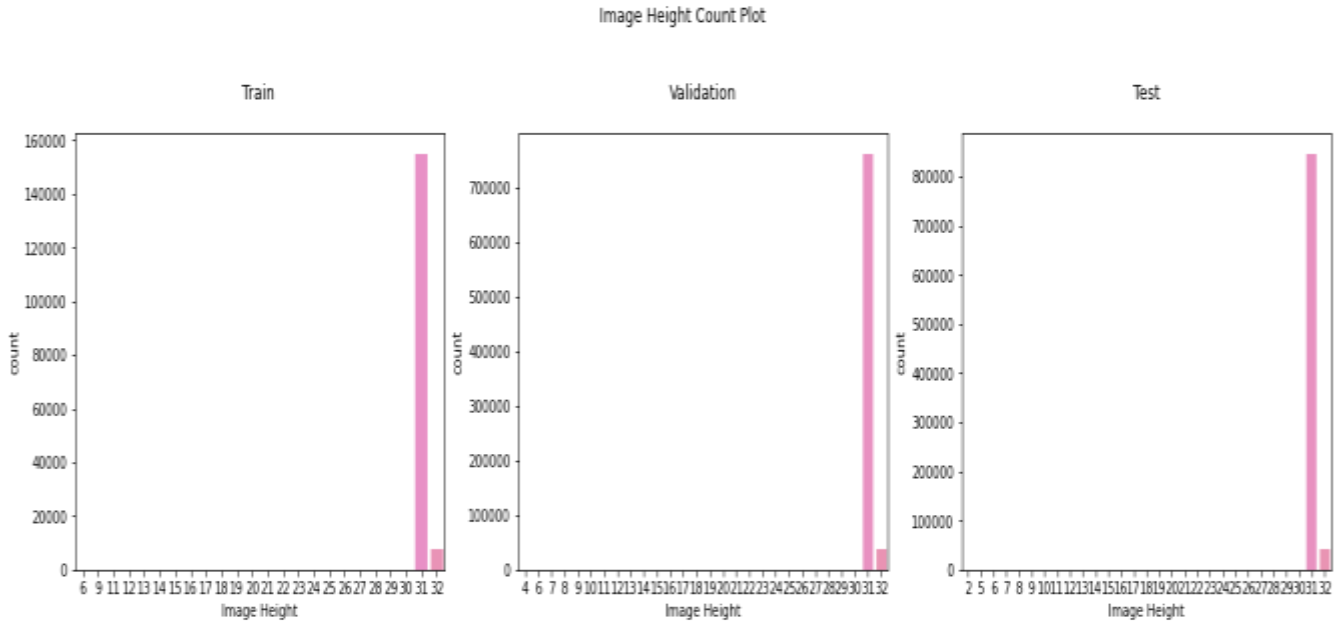


Figure 3.5: Count Plot for Image Height.

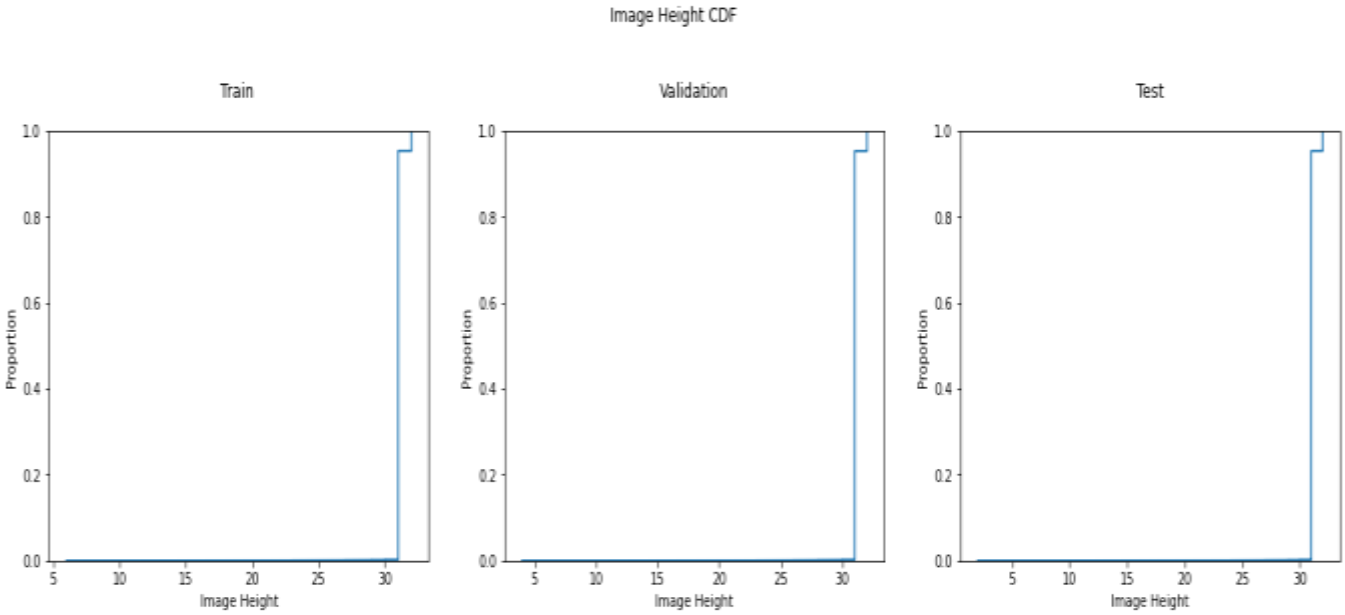


Figure 3.6: CDF for Image Height.

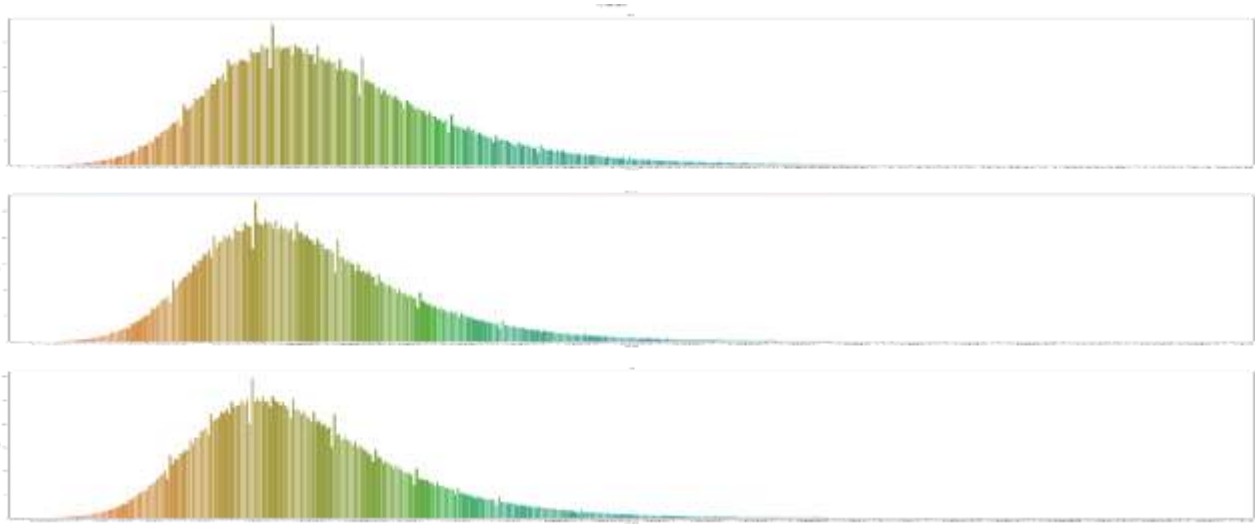


Figure 3.7: Count Plot for Image Width**.**

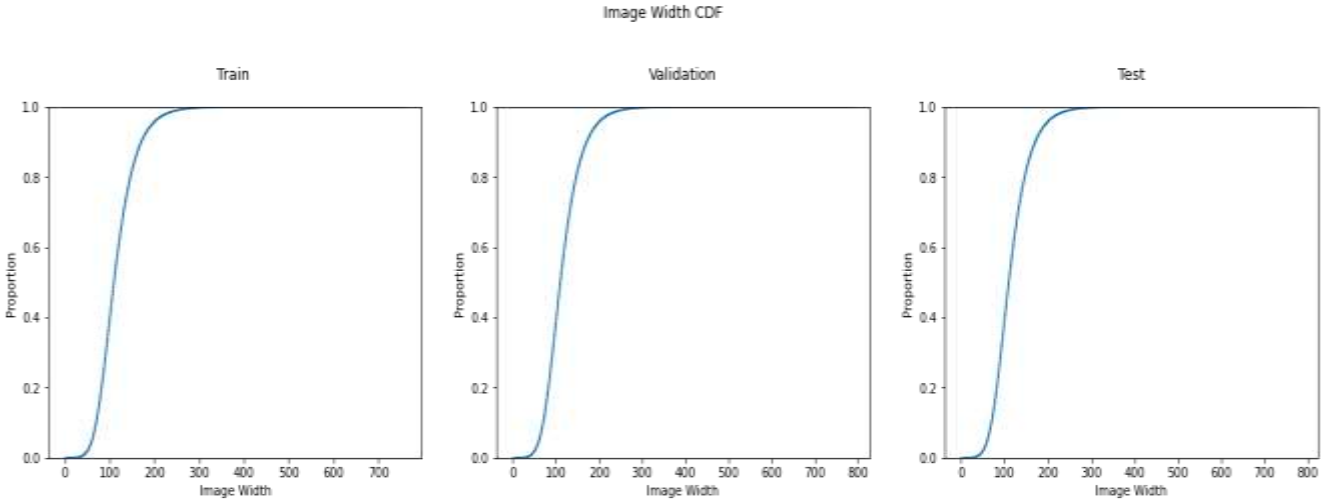


Figure 3.8: CDF for Image Width

3.5.3. Model Structure

The recognition model has two input layers and one output layer. A prediction model was built from the originally generated training model. The training model consists of four blocks of CNN. The first three blocks of CNN has three CNN and maintain the same filters. While the last block has 2 CNN connected to a batchnormalizer. The image is firstly passed in convolutional layer three times. This layer uses set of detectable features to be applied on the filters after which max pooling layer is used to reduce the spatial size of the image as shown in Table 3.1. The Feature extracted from CNN was further passed to batchnormalizer before feeding to RNN to prevent over fitting. The output obtained from RNN after 3 times BiLSTM is passed to Dence Layer where CTC is applied to specify a function mapping from the given filters to a vector. Finally, the extracted and decoded text from CTC is converted to speech using Google text to speech and the output is communicate to the VIP in audio format.

Total parameters used for the training are, 27,529,919. 2752799 of which were trainable while 1920 parameters were not trainable. The model structure can be summarized as follows:

The proposed systems CRNN architecture as shown in figure 3.2 has the following properties:

1. Eleven (11) Convolutional Layers split in four (4) blocks. Each block is followed by Batch Normalizer and Max Pooling layers;
2. A Spatial Transformation Layer which is integrated into the CRNN to increase spatial variation in an efficient manner. It achieves this by cropping, transforming, and scaling the region of interest in the given sample.
3. Three (3) Bidirectional LSTM layers for both forward and backward feature extraction.
4. The CTC layer which selects the best possible output from the arrays of output using dynamic programming
5. The text-to speech layer which convert the recognized word to the VIP in audio form.

Table 3.4: Summary of Training Model **Model: "oculus_crnn_v1"**

Layer (type)	Output Shape	Param #	Connected to
image (InputLayer)	[(None, 200, 50, 1)]	0	
block_1_conv_1 (Conv2D)	(None, 200, 50, 64)	640	image[0][0]
block_1_conv_2 (Conv2D)	(None, 200, 50, 64)	36928	block_1_conv_1[0][0]
block_1_conv_3 (Conv2D)	(None, 200, 50, 64)	36928	block_1_conv_2[0][0]
bn_1 (BatchNormalization)	(None, 200, 50, 64)	256	block_1_conv_3[0][0]
maxpool_1 (MaxPooling2D)	(None, 100, 25, 64)	0	bn_1[0][0]
block_2_conv_1 (Conv2D)	(None, 100, 25, 128)	73856	maxpool_1[0][0]
block_2_conv_2 (Conv2D)	(None, 100, 25, 128)	147584	block_2_conv_1[0][0]
block_2_conv_3 (Conv2D)	(None, 100, 25, 128)	147584	block_2_conv_2[0][0]
bn_2 (BatchNormalization)	(None, 100, 25, 128)	512	block_2_conv_3[0][0]
maxpool_2 (MaxPooling2D)	(None, 50, 12, 128)	0	bn_2[0][0]
block_3_conv_1 (Conv2D)	(None, 50, 12, 256)	295168	maxpool_2[0][0]
block_3_conv_2 (Conv2D)	(None, 50, 12, 256)	590080	block_3_conv_1[0][0]
block_3_conv_3 (Conv2D)	(None, 50, 12, 256)	590080	block_3_conv_2[0][0]
bn_3 (BatchNormalization)	(None, 50, 12, 256)	1024	block_3_conv_3[0][0]

maxpool_3 (MaxPooling2D)	(None, 25, 6, 256)	0	bn_3[0][0]
block_4_conv_1 (Conv2D)	(None, 25, 6, 512)	1180160	maxpool_3[0][0]
block_4_conv_2 (Conv2D)	(None, 25, 6, 512)	2359808	block_4_conv_1[0][0]
bn_4 (BatchNormalization)	(None, 25, 6, 512)	2048	block_4_conv_2[0][0]
reshape (Reshape)	(None, 25, 3072)	0	bn_4[0][0]
fc_9 (Dense)	(None, 25, 512)	1573376	reshape[0][0]
dropout (Dropout)	(None, 25, 512)	0	fc_9[0][0]
lstm_10 (LSTM)	(None, 25, 1024)	6295552	dropout[0][0]
lstm_10_back (LSTM)	(None, 25, 1024)	6295552	dropout[0][0]
add_26 (Add)	(None, 25, 1024)	0	lstm_10[0][0] lstm_10_back[0][0]
lstm_11 (LSTM)	(None, 25, 512)	3147776	add_26[0][0]
lstm_11_back (LSTM)	(None, 25, 512)	3147776	add_26[0][0]
add_27 (Add)	(None, 25, 512)	0	lstm_11[0][0] lstm_11_back[0][0]
lstm_12 (LSTM)	(None, 25, 256)	787456	add_27[0][0]
lstm_12_back (LSTM)	(None, 25, 256)	787456	add_27[0][0]
concatenate_13 (Concatenate)	(None, 25, 512)	0	lstm_12[0][0] lstm_12_back[0][0]
label (InputLayer)	[(None, None)]	0	
fc_12 (Dense)	(None, 25, 63)	32319	concatenate_13[0][0]
ctc_loss (CTCLayer)	(None, 25, 63)	0	label[0][0] fc_12[0][0]

Out of 27,529,919 Parameters that were used for the training, 27,527,999 were trainable while 1,920 were not trainable. Table 3.4 shows the summary the different blocks of convolutions and other activities that took place during the model training.

3.5.4. Loss Function

For the purpose of this work, CTC loss function was used. CTC loss is very helpful in text recognition problems. It helps to get the best possible output from the arrays of outputs sent into the dense layer.

Hence, from the trained model the training and validation loss as shown in Figure 3.9 indicates the rise of accuracy with increase in number of epochs. While the Figure 3.10 indicates the decrease in training and validation loss as the number of epochs increases.

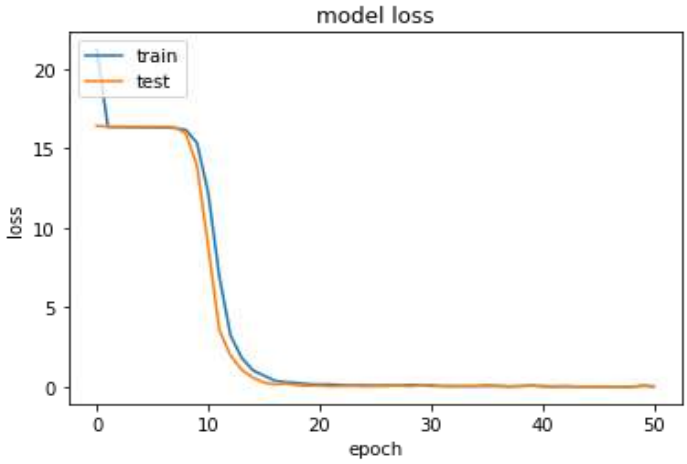


Figure 3. 9: Graph of model loss versus epochs

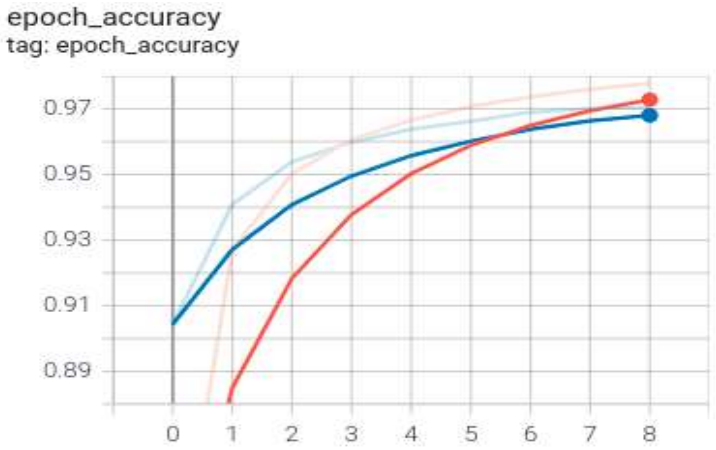


Figure 3.10: Epoch_accuracy

From the above plots, the following inferences were made:

i. **Observation(s):**

1. Images have height ranging between 2 and 32.
2. Almost 99% of the Images have a height of 31.
3. Image Width varies between 1 and 799.
4. Majority of the Images have a width between 30 and 190.

ii. Sampling Dataset

The sampling will be performed as follows:

a. Small images with width and/or length less that 5pixels will not be considered in the sampled dataset.
b. 100, 000 images from the original train dataset will be sampled.
c. 120, 000 will be sampled from the original 120, 000 validation dataset.
d. 80, 000 images will be sampled from the original test dataset.

```
train = train.sample(100000, random_state=3).reset_index(drop=True)

validation = validation.sample(120000, random_state=3).reset_index(drop=True)

test = test.sample(80000, random_state=3).reset_index(drop=True)
```

3.5.5. Model Training

For model training Keras callbacks functionality was used to save the weights of the best model on the basis of validation loss.

3.5.6. Preparing Data for our Recognition Neural Net

Outlined next are the steps required to set parameters for our Neural net

```
#1. Set Parameters for our images

batch_size = 16
img_height = 50
img_width = 200
max_label_length = 25
alphabets = set(string.ascii_letters+string.digits)
downsample_factor = 4
epochs = 100
early_stopping_patience = 10
early_stopping = keras.callbacks.EarlyStopping(
   monitor="val_loss", patience=early_stopping_patience, restore_best_weights=T
rue
```

```
)

#2. Fetch image from Data file:

train_df = pd.read_csv('Data/train.csv', na_filter=False)
validation_df = pd.read_csv('Data/validation.csv', na_filter=False)
test_df = pd.read_csv('Data/test.csv', na_filter=False)

#3. Create image and label variable for train, validation and test datasets:

#images
x_train = train_df.copy()
x_val = validation_df.copy()
x_test = test_df.copy()

#labels
y_train = x_train.pop('label')
y_val = x_val.pop('label')
y_test = x_test.pop('label')

#4. Create character encoder-decoder function:

# Mapping characters to integers
char_to_num = layers.StringLookup(
    vocabulary=list(alphabets), mask_token=None
)

# Mapping integers back to original characters
num_to_char = layers.StringLookup(
    vocabulary=char_to_num.get_vocabulary(), mask_token=None, invert=True
)

def encoder_decoder(img_path, label):
    # 1. Read image
    img = tf.io.read_file(img_path)
    # 2. Decode and convert to grayscale
    img = tf.io.decode_png(img, channels=1)
    # 3. Convert to float32 in [0, 1] range
    img = tf.image.convert_image_dtype(img, tf.float32)
    # 4. Resize to the desired size
```

```
    img = tf.image.resize(img, [img_height, img_width])
    # 5. Transpose the image because we want the time
    # dimension to correspond to the width of the image.
    img = tf.transpose(img, perm=[1, 0, 2])
    # 6. Map the characters in label to numbers
    label = char_to_num(tf.strings.unicode_split(label, input_encoding="UTF-8"))
    # 7. Return a dict as our model is expecting two inputs
    return {"image": img, "label": label}
```

#5. Preprocess dataset for neural net by adding encoder-decoder, batch_size and buffer prefetch:

```
train_dataset = tf.data.Dataset.from_tensor_slices((x_train, y_train))

train_dataset = (
    train_dataset.map( encode_single_sample, num_parallel_calls=tf.data.AUTOTU
NE)
    .batch(batch_size)
    .prefetch(buffer_size=tf.data.AUTOTUNE)
)

validation_dataset = tf.data.Dataset.from_tensor_slices((x_val, y_val))
validation_dataset = (
    validation_dataset.map(
        encode_single_sample, num_parallel_calls=tf.data.AUTOTUNE
    )
    .batch(batch_size)
    .prefetch(buffer_size=tf.data.AUTOTUNE)
```

3.5.7. Algorithm for Text Recognition

a. we create a function to build CRNN model, with the following parameters:

height: The height of cropped images.

width: The width of cropped images.

color: Whether the inputs should be in color (RGB).

filters: The number of filters to use for each of the 7 convolutional layers.

rnn_units: The number of units for each of the RNN layers.

dropout: The dropout to use for the final layer.

rnn_steps_to_discard: The number of initial RNN steps to discard.

pool_size: The size of the pooling steps.

stn: Whether to add a Spatial Transformer layer.

b. create 4 blocks with 3 CNN layers
c. apply spatial transformation
d. create 3 bidirectional LSTM layers
e. create backbone
f. add dropout layer
g. add output neuron with filter = length of alphabet + 1, and softmax activation function
h. final model
i. create prediction model
j. create loss function using CTC loss
k. create training model, using model(h) input, labels, input_length and label_length as inputs, and loss as output for the CTC Decoder
l. create text recognition class with an init function to initialize model built using the build_model function

3.6. Use case Diagram

The use case diagram of the proposed system is presented in Figure 3.10. Since the aim of the system is to enable the visually impaired person access print document and scene images, the user will only have to interact with the system at the input level and output level. At the input level, the user captures scene images or print documents with an attached Camera device and transfers to the user machine through wireless device. The users Machine accept the image and carry out all the necessary preprocessing and transformation. At the end of the transformation, the

user gets audio feedback of the transformed text via the output module of the system.

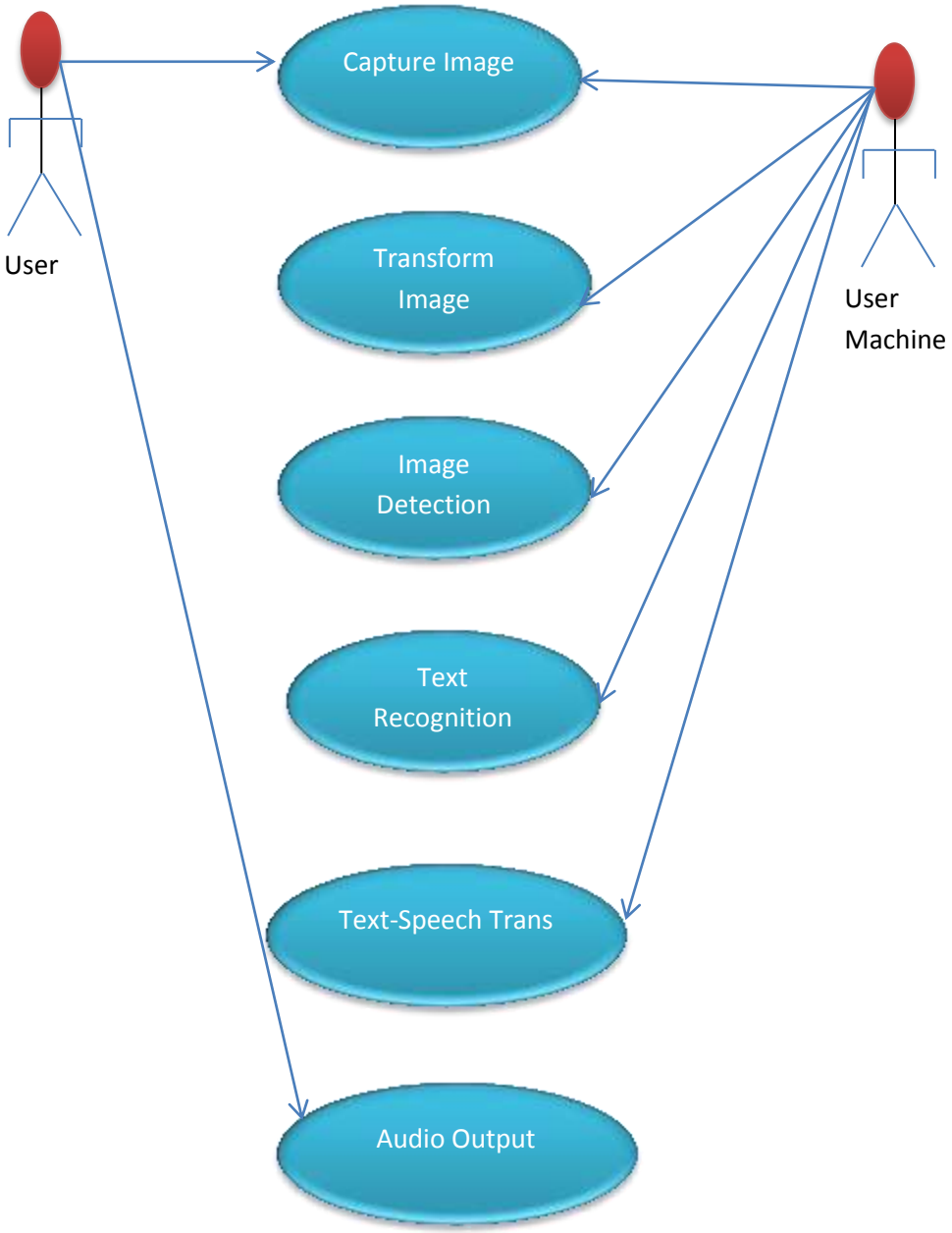


Figure 3.11: Use Case Diagram of the Proposed System.

3.7. System Flowchart

The system as earlier said has three major modules, Image capturing and pre-processing, text localization and detection, finally feature extraction and text detection module. Figure 3.12 describes the flow of activities of the entire system modules.

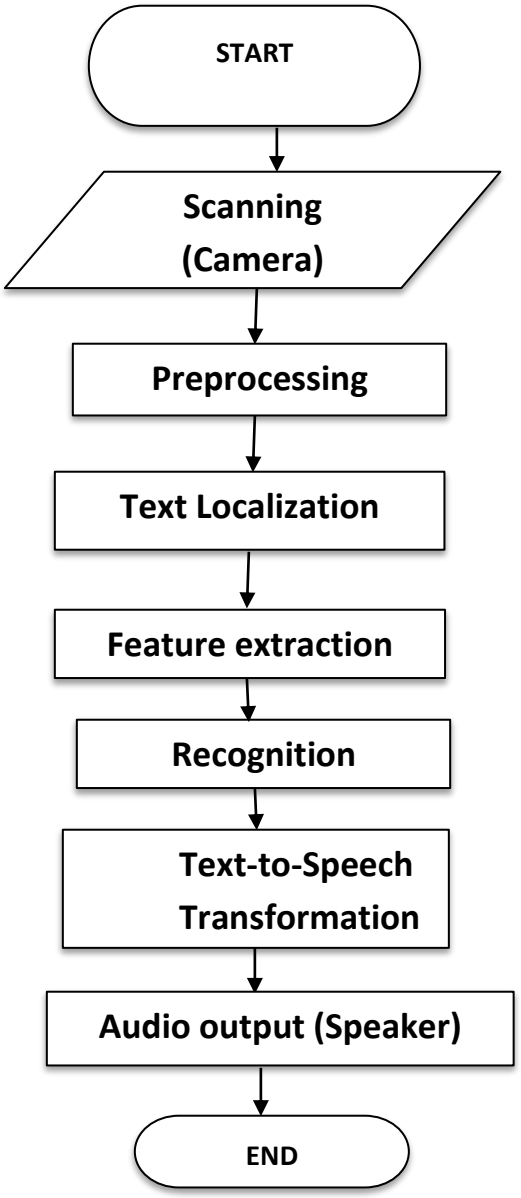


Figure 3.12: Flowchart of the Proposed System

3.8 Features of the Proposed System

1. The system connects once it detects a camera.
2. The Camera captures the user screen, scene images or any print document via a wearable camera or an attached camera and submits to user Machine which uses OCR technology to translate the captured image to text then to speech and finally voiced it out to the user through earphone or system speaker.
3. The system runs automatically as soon as the user system is turn on.
4. The system transmits every output to the visually impaired person in an audio format.
5. The system is purely offline based. it requires no internet connection to function; hence, the fear of network failures and poor internet services are completely eradicated
6. The system supports reading of documents and scene images.
7. The system is a real time application
8. It guarantees a comfortable working environment for the VIP
9. All the processing is done within the User Machine. Hence, The system is less expensive
10. The system is user friendly, it allows effective interaction between the user and the machine
11. The system enables the VIPs to work effectively without any assistance from an external person.

3.9 Expectations of the Proposed System

The proposed system is expected to:

1. Enhance employment chances for the VIP

2. Allow the blind user to read print document with both multipart and complex background
3. Allow the user to read the screen content
4. Allow the user to access scene images text
5. Operate in zero internet services and network failure
6. Reduce the high rate of job-loss associated with vision impairment
7. Run effectively with trivial human intervention

3.10 User Machine System Requirements Specification

The system shall have the following functional and non-functional requirement specifications:

3.10.1 Functional Requirements

1. The Camera shall capture the user screen or any print document and submits to user Machine which uses OCR technology to translate the captured image to text then to speech and finally voiced it out to the user through earphone.
2. The system shall be enabled to run automatically as soon as the user system is turned on.
3. The System shall have machine learning capabilities for cross-matching the captured images.
4. The system shall terminate only when the system is turned off otherwise, it keeps running.
5. The system shall have a wireless earphone for audio display.

3.10.2 Non-functional Requirements

1. The system shall support more than 5000 transactions per day

2. The wearable wireless camera shall have a configuration of not less than 32 pixels
3. The camera should support auto focusing
4. The camera should have a 360 degree rotating capability
5. The system shall support a peak transaction rate of 10seconds per transaction.
6. The system shall be offline base to enable 24hours transaction
7. The system shall use a light weight server less database such as SQLite for easy access

3.11 Software Development Tools

Development of A Camera-Based Text detection and recognition system is purely an AI system. Hence, for an improved performance, Python programming language which was proven to be one of the best software development tools for implementing Computer Vision problems was adopted. The following software development tools were used:

PyQt5: used for designing the graphic user interface (GUI) of the developed system. It is one of the commonly used python modules for building GUI applications. This is largely use due to its simplicity during usage. All that is usually required is to drag necessary widgets to develop a form.

MATPLOTLIB: is a data visualization and graphical plotting library for Python Programming language. It offers a viable open source alternative to MATLAB. Also, it uses an object oriented API to embed plots in Python applications.

NUMPY: is an Acronym for Numerical Python. It is a library consisting of multidimensional array objects and a collection of routines for processing those

arrays. Mathematical and logical operations on arrays are handled efficiently using NUMPY.

SciPy: is a free and open-source Python library used for scientific computing and technical computing. It adds significant power to the interactive Python session by providing the user with high-level commands and classes for manipulating and visualizing data.

TesorFow : is the machine learning framework adopted for the system. It has a comprehensive, flexible ecosystem of tools, libraries and community. While Keras **Keras**: used for the back end development. It is an open source software library that provides a python interface for artificial neural net.

Google Colab: was adopted for the development environment. It is a free platform from Google that allows users to code in python. It uses graphical Processing Unit (GPU).

3.12 Software Specifications

The minimum software requirements for the system are:

i. Windows XP Operating System

ii. Standard windows explorer with active antivirus software

3.13 Hardware Specifications

The minimum hardware requirements for the system are:

i. Laptop or desktop computer with keyboard and mouse

ii. 8GB Random Access Memory

iii. 64-bit operating system, x64-based processor

iv. ROM: 300GB

v. 300 GB Hard Disk Drive

vi. UPS Unit for power backup

CHAPTER FOUR

RESULTS AND DISCUSSION

The recognition model developed in chapter 3 was evaluated on test data, evaluation matrices and human subjects. This s to ascertain to what extend the system was to meet up with user's requirements.

4.1 **Evaluation on Test Data**

The proposed system's model was evaluated on test data as shown in figures 4.1a-4.1p; both the loss and accuracy were recorded as follows:

Loss on Test Data: 0.02456

Accuracy on Test Data: 98%

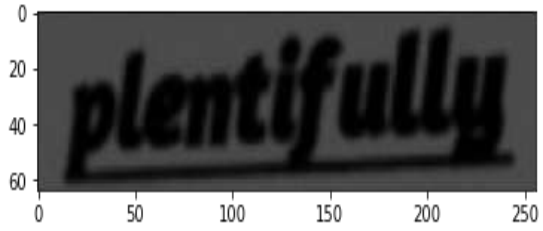


Figure 4.1a: Test data1 **Original: PLENTIFULLY**
Predicted: PLENTIFULLY

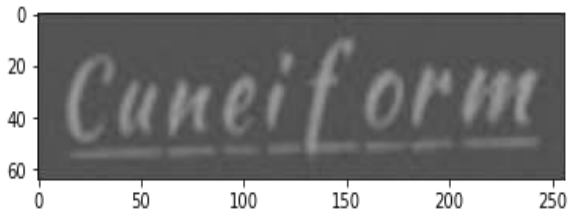


Figure 4.1b: Test data2 **Original: CUNEIFORM**
Predicted: CUNEIFORM

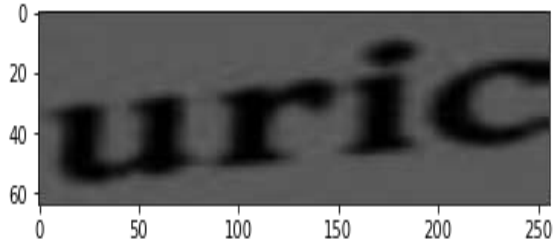


Figure 4.1c: Test data3 **Original: URIC**
Predicted: URIC

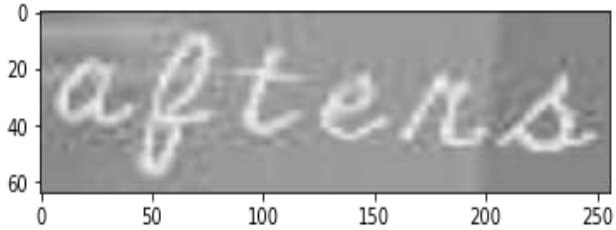


Figure 4.1d: Test data4 **Original: AFTERS**
predicted: AFTERS

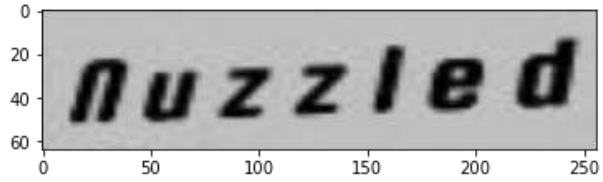


Figure 4.1e: Test data5 Original: NUZZLED
Predicted: NUZZLED

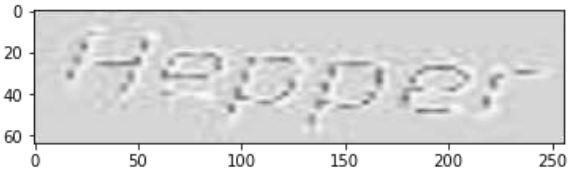


Figure 4.1f: Test data6 Original: HEPPER
Predicted: HEPPE

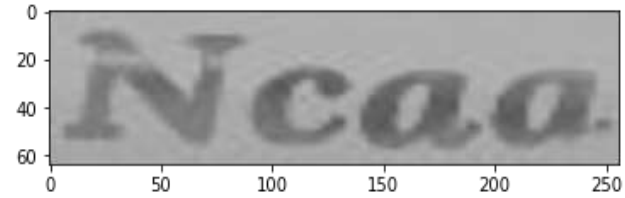


Figure 4.1g: Test data7 Original: NCAA
Predicted: NCAA

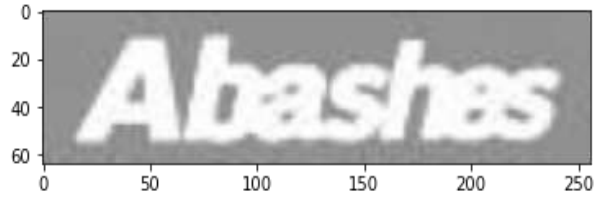


Figure 4.1h: Test data8 Original: ABASHES
Predicted: ABASHES

Figure 4.1i: Test data 9 Original: WINGDINGS
Predicted: WINGDINGS

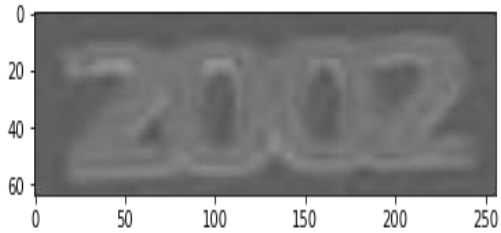


figure 4.1j: Test data10 Original: 2002
Predicted: Z002

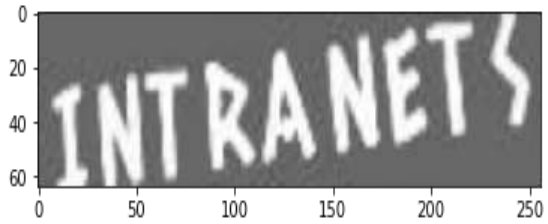


Figure 4.1k: Test data10 Original: INTRANETS
Predicted: INTRANET2

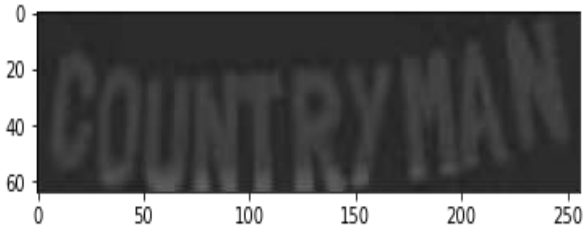
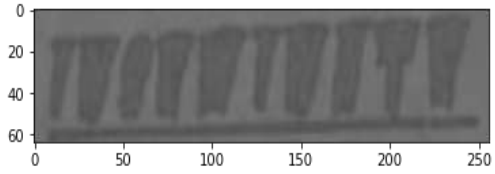


figure 4.1l: Test data 11 Original: COUNTRYMAN
Predicted: COUNTRYMAN



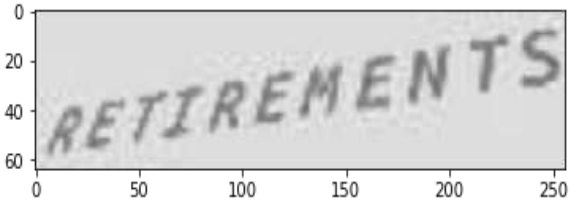


Figure 4.1m: Test data12 Original: INSEMINATE Predicted: TARRIAIIA

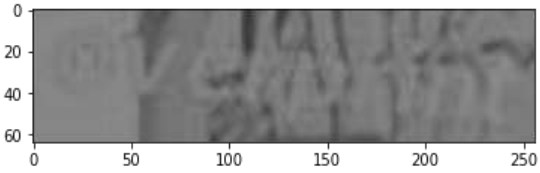


figure 4.1n: Test Data13 Original: RETIREMENTS Predicted: RETTREMENTS

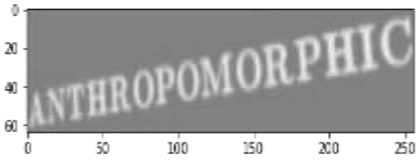


Figure 4.1o: Test Data14 **Original: OVERPRINT Predicted: OVERRIIT**

figure 4.1p: Test data15 Original: ANTHROPOMORPHIC Predicted: INTEHOOPOPPHI

4.2 **Evaluation Matrices**

This evaluation was based on the following evaluation matrices Precision, Recall and f-Measure

Precision = (No. of text in input image ∩ No. of detected text in output image) ÷ No. of detected text in output image. i.e. **precision = (NTI ∩ NDTO)/ NDTO**

Recall = (No. of text in input image ∩ No. of detected text in output image) ÷ No. of text in input image**. i.e. Recall = (NTI ∩ NDTO)/ NTI**

f-measure = 2 (Precision × Recall) ÷ (Precision + Recall)

Where: NTI = No. of text in input image, **NDTO =** No. of detected text in output image.

After evaluating the system with existing data sets as illustrated in figures 4.1a to figure 4.1p above, the following deductions based on Precisions, Recall and f-measures were made. When 10k images were considered with condition that every character to be exactly correct, the results were recorded as follows:

Precision Score Metrics

```
[71] #Multilabel Precision score for 10 batches of validation data:

     precision_sc = precision_score(orig_texts, pred_texts, average=None)
```

```
precision_sc.ravel()

array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.])
```

Confusion Matrix

```
[73] cm = confusion_matrix(orig_texts, pred_texts)

[74] cm

     array([[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
            [0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
            [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
            [0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
            [0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
            [0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0],
            [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0],
            [0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0],
            [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0],
            [0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0],
            [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0],
            [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0],
            [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0],
            [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0],
            [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]])

[79] f1_s = f1_score(orig_texts, pred_texts, average=None)

[80] f1_s

     array([1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1.])
```

R

1.])

The result obtained from the recognition model was compared with that of the existing systems that are very close to our system and the result is displayed in figure

Table 4.1: Comparing proposed system's Result with Existing System

	Accuracy,	Precision(10 batches)	Recall	f-measure	Approach/Data Set

Our result	98%	1	1	1	CRAFT and CRNN. Synth90k
Shikha *et al.*, (2020)	74.24%	0.57	0.63	74.24%	Otsu's thresholding And CRNN: Ancient scripts
Kambala *et al.*, (2020)	59.9%				VGG Architecture and CRNN, MJSynth" dataset.
(Uma et al., 2018)	85.12%	84.66%	85.59%	85.12%	MSER, ICDAR 2011

4.3 Usability Testing and Evaluation Using Human Subjects

The system was tested with two different groups of participants. Each of the groups consists of five test participants. The first group of participants has a maximum age of twenty five (25) years, while the second group of participants consists of participants with a minimum age of thirty (30) years. The participants have been grouped into these categories to investigate if the acceptability of the system is tied to age differences. Figures 4.2 and 4.3 shows the graphs of perceived usefulness and ease of use for group one, while figures 4.4 and 4.5 shows the graph of perceived usefulness and ease of use for group one accordingly.

Perceived Usefulness

Participants	Perceived usefulness	Perceived ease of use

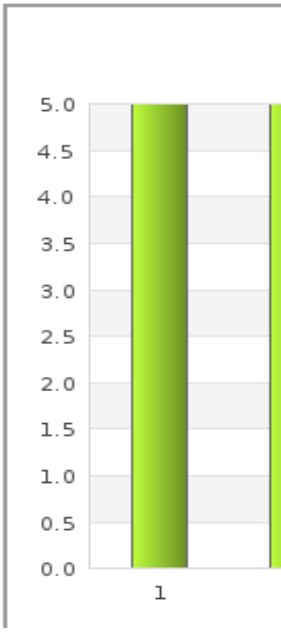


Figure 4.2: Graph of Perceive Usefulness for Group 1

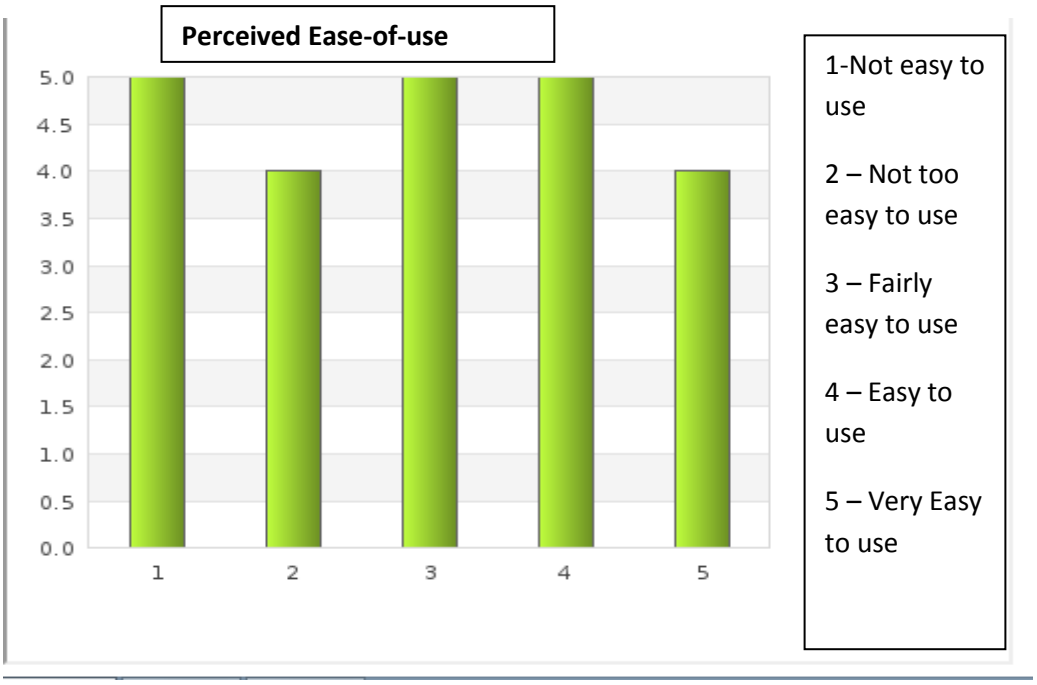


Figure 4.3: Graph of Perceived ease of use for group 1

Table 4.2: Rating Result For group 1(users with maximum age of 25)

Part1	5	5
Part2	5	4
Part3	5	5
Part4	5	5
Part5	5	4

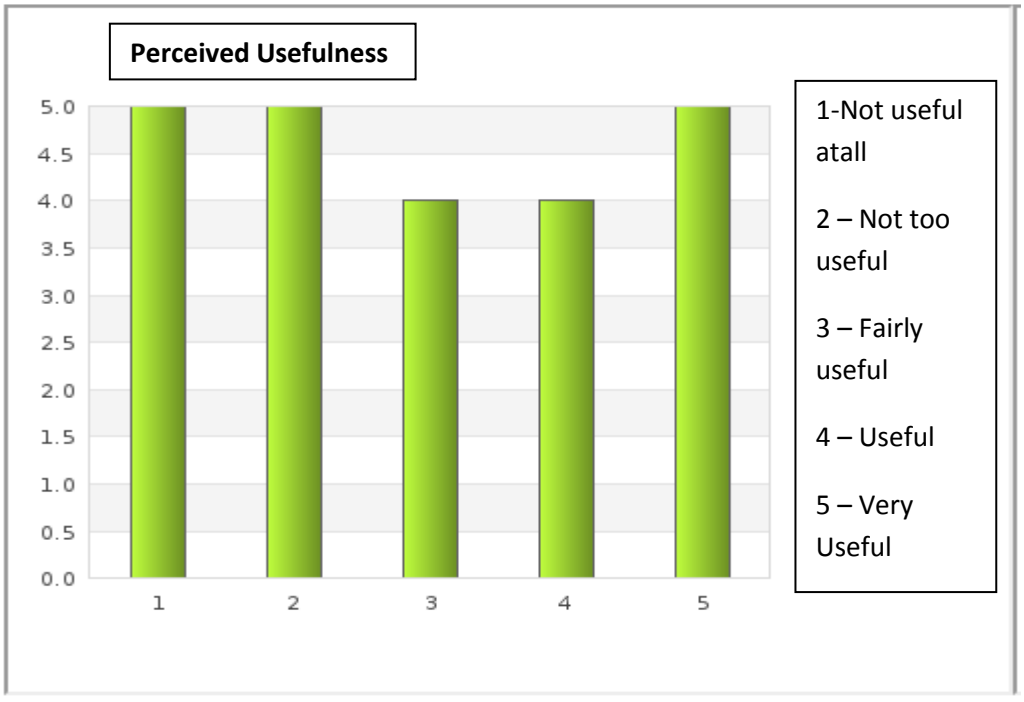


Figure 4.4: Graph of Perceive usefulness for group 2

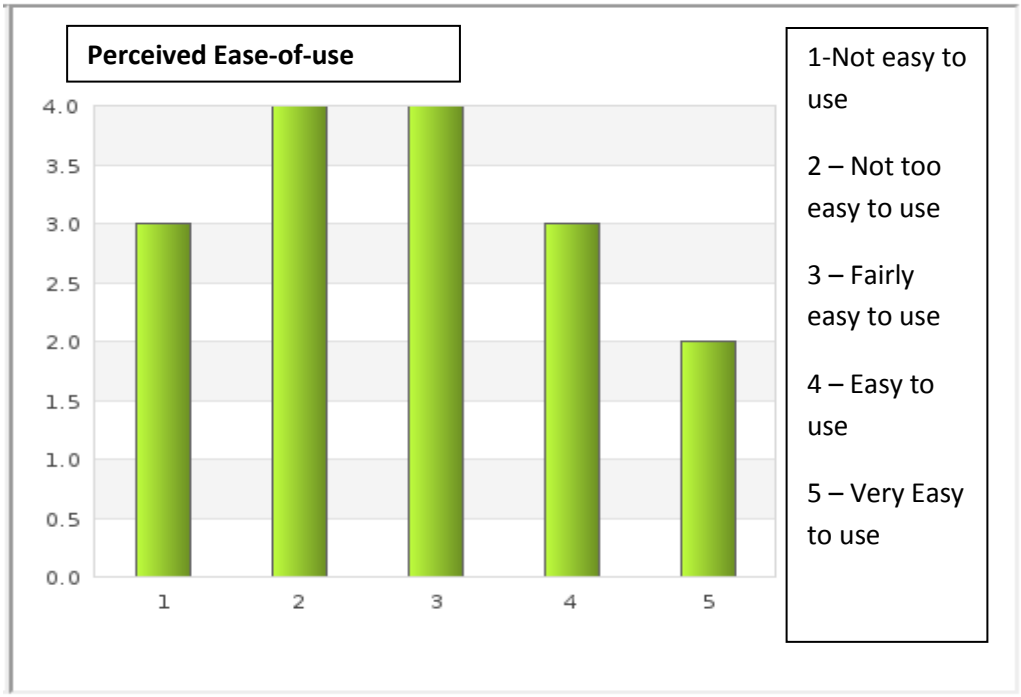


Figure 4.5: Graph of Perceived Ease-of-Use for Group 2.

Table 4.3: Rating Result for Group 2(users with minimum age of 30)

Participants	Perceived usefulness	Perceived ease of use
Pat1	5	3
Part2	5	4
Part3	4	4
Part4	4	3
Part5	5	2

From figures 4.1, 4.2, 4.3 and 4.4 above, it was observed that the first group of test participants with a maximum age of twenty five (25) rated the system highly both in terms of perceived usefulness and perceived ease-of-use. The second group of

participants also rated the system highly with respect to usefulness, but fairly high with respect to ease of use.

It can be inferred from this result that while this system might be perceived as useful for all classes of individuals, there might be need to consider some other elements in order to make this system easy to use across age ranges.

This finding is however not a disadvantage since our system is targeted at users within the age ranges of the first group.

4.3.1 Usability Result Analysis

After testing and evaluating the system, the results obtained from rating the systems were further analyzed using Statistical Analysis Software (SAS). This was done using Two Sample Paired t-test for the Means of Perceived Usefulness and ease of use for users with maximum age of 25 and those with minimum age of 30 as shown in tables 4.2 and 4.3 above.

4.3.2 Result Analysis Based On Perceived Usefulness

Two Sample Paired t-test for the Means of Perceived Usefulness of users with maximum age of 25 (PU25) and users with minimum age of 30 (PU30)

Sample Statistics

Group	N	Mean	Std. Dev.	Std. Error
PU_25_	5	5	0	0
PU_30_	5	4.6	0.5477	0.2449

Hypothesis Test

Null hypothesis: Mean of (PU_25_ - PU_30_) = 0

Alternative: Mean of (PU_25_ - PU_30_) ^= 0

t Statistic	Df	Prob > t
1.633	4	0.1778

95% Confidence Interval for the Difference between Two Paired Means

Lower Limit	Upper Limit
-0.28	1.08

The *t* statistic is 1.633 and the associated *p*-value (0.1778). Since the p-value (0.1778) is greater than 0.05 (5 % level of significance), i.e. $p > \alpha$ $(0.1778 > 0.05)$, we do not reject the null hypothesis. Hence, the p –value of 0.1778 provides evidence at the 5% level of significance that there is no difference in perceived usefulness of the system among the users with maximum age of 25 and those with minimum age of 30. The confidence interval indicates that you can be 95% confident that there is no difference in perceived usefulness of the system among the users with maximum age of 25 and those with minimum age of 30. Fgures 4.5 shows the MEANS of t distribution plot of PU25 and PU30

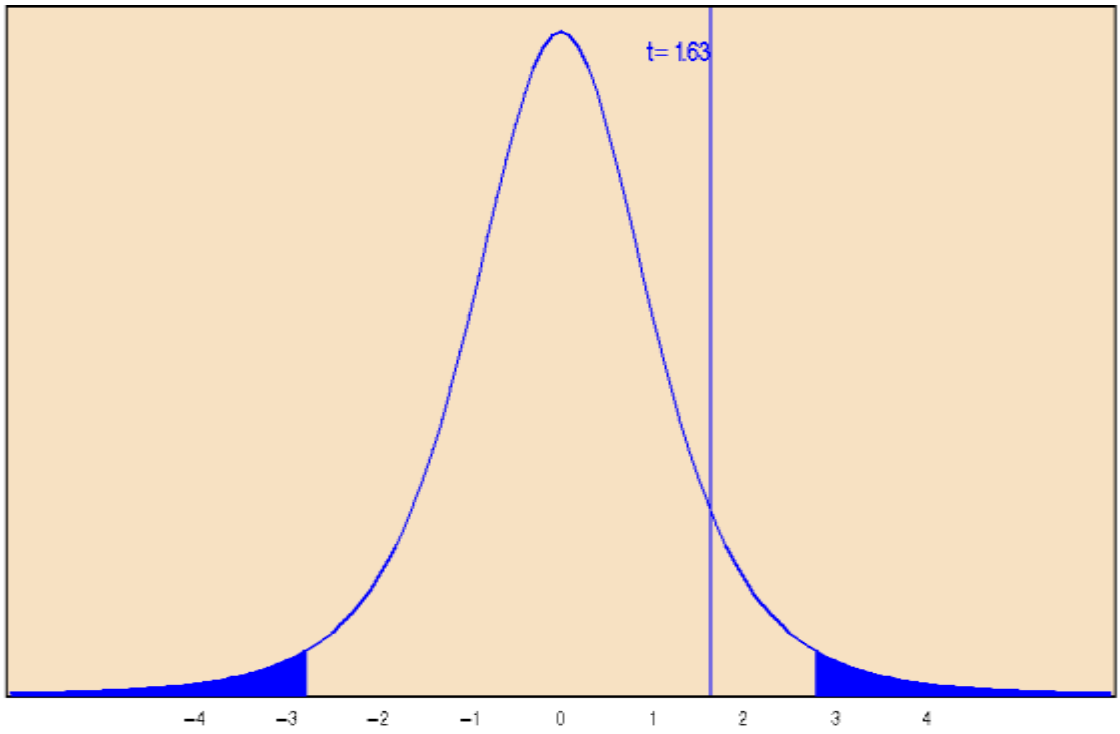


Figure 4.6: t distribution plot of PU25 and PU30 MEANS

As shown in the graph of t distribution plots in figure 4.6, the t value of 1.633 falls within the acceptable region. Hence, we do not reject the null hypothesis.

This also implies that there is no difference in perceived usefulness of the system among the users with maximum age of 25 and those with minimum age of 30.

4.3.3 Analysis of the Result Based On Perceived Ease of Use

Two Sample Paired t-test for the Means of Perceived Ease of Usefulness for Users with maximum age of 25(PEU_25) and people with minimum age of 30 (PEU_30)

Sample Statistics

Group	N	Mean	Std. Dev.	Std. Error
PEU_25_	5	4.6	0.5477	0.2449
PEU_30_	5	3.2	0.8367	0.3742

Hypothesis Test

Null hypothesis: Mean of (PEU_25_ - PEU_30_) = 0
Alternative: Mean of (PEU_25_ - PEU_30_) ^= 0

t Statistic	Df	Prob > t
3.500	4	0.0249

95% Confidence Interval for the Difference between Two Paired Means

Lower Limit	Upper Limit
0.29	2.51

The *t* statistic is 3.500 and the associated *p*-value (0.0249). Since the p-value (0.0249) is less than 0.05 (5 % level of significance), i.e. $p < \alpha$ (0.0249< 0.05), we reject the null hypothesis. Hence, the p –value of 0.0249 do not provides enough evidence at the 5% level of significance that there is no difference in perceived usefulness of the system among the users with maximum age of 25 and those with minimum age of 30. The confidence interval indicates that you can be 95% confident that there is difference in perceived usefulness of the system among the users with maximum age of 25 and those with minimum age of 30.

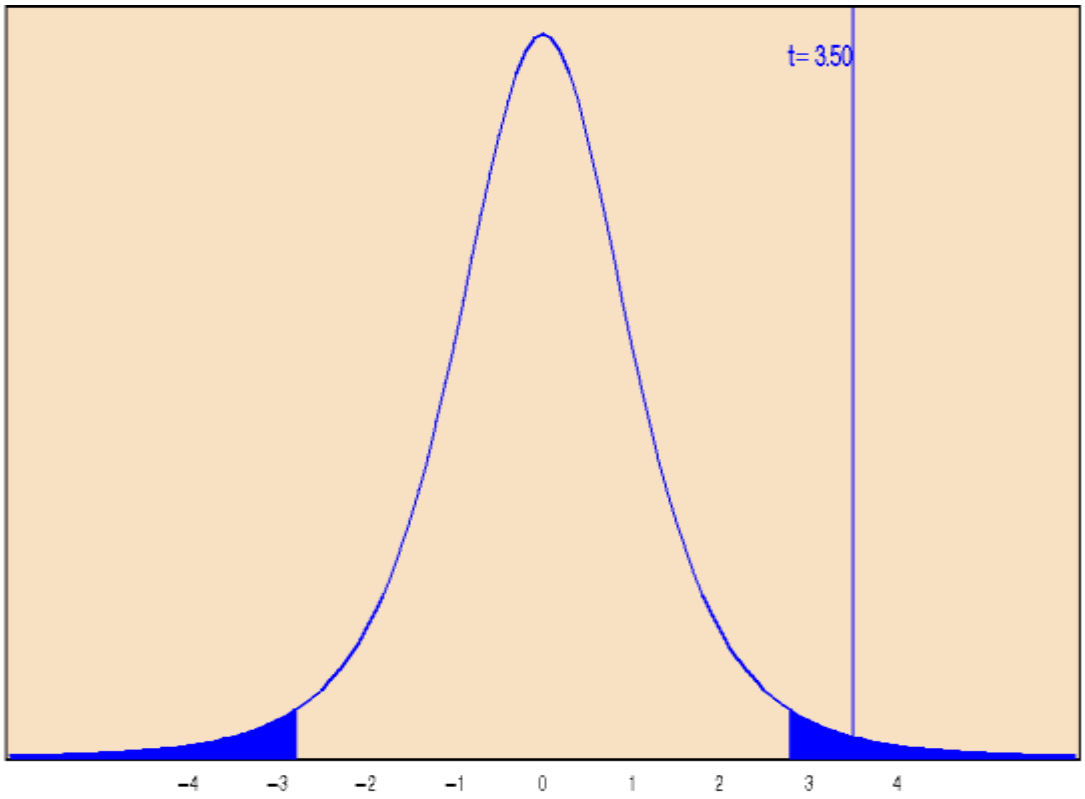


Figure 4.6: t distribution plot of PEU25 AND PEU30 MEANS

As shown in the graph of t distribution plots in figure 4.6 , the t value of 3.500 falls outside the acceptable region. Hence, we reject the null hypothesis.

This also implies that there is difference in perceived usefulness of the system among the users with maximum age of 25 and those with minimum age of 30

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.4. Conclusion

The goal of this study which was to develop a camera-based scene and document text reading system for the visually impaired by leveraging the functionalities of OCR for text recognition was met. Having critically analyzed some existing system and their methods, this system was successfully implemented using deep learning approach, precisely CRAFT for Text detection and CRNN which combines the functionalities of CNN, RNN and CTC loss for an optimal Character Recognition. For an improved performance, Synth90k synthetic text dataset provided by the Visual Geometry Group (VGG) architecture was explored, analyzed and enhanced along with CRAFT Architecture which is suitable for detecting Curved images. *Synth90k* data set contains 90 million images of 90k unique English words. The text recognition accuracy was obtained to be 98% considering images with condition that every character to be exactly correct. This application was deployed for real time which is in line with state of the art result.

Though, there are so many existing assistive devices for the blind, so many researchers had also made frantic effort to ensuring independent living of the concerned group. The truth still remains that none of these existing systems have been able to totally eradicate the barriers confronting the visually impaired. Secondly, the very wonderful devices among the existing assistive technology are either not accessibly or affordable by those concerned. Hence, the need for the development of an affordable and accessible Robust Camera-Based Text recognition system for the visually impaired.

5.5. Recommendations

Aside bringing to light the state-of-the-art of text detection and recognition full concentration on deep learning based approaches, there are still some research problems that have been identified and reported. This study leverages the functionalities of CRAFT architecture for text localization and detection. Further work should be extended to creating text detector from Scratch. The training data set used was the one provided by VGG, further studies should work toward providing the training data set. The scope of this study is limited to capturing print documents, scene images and images from complex background. Further research should incorporate capturing images on moving objects. This field of study is a very broad area that has gained thousands of researcher's interest. Hence, an endless development and improvement on both application-wise and research-wise remains undisputable.

5.6. Contribution to Knowledge

Computer Vision is a field in Artificial Intelligence which remains a research area with ever increasing attention. It remains one of the top technologies that enable the interaction between the digital world and physical world. Its contribution to ensuring independent living of the visually impaired cannot be overemphasized and Camera-Based Optical Character Recognition in terms of performance remains one of the applications with high demand. Hence, since this studies focused on deep learning approach of text detection and recognition. Major contributions are:

i. Built more efficient text recognition model

ii. Improved CTC to return a better adapted text to speech.

iii. Integrated CRAFT detector into an OCR pipeline for a better text to speech output.

References

Abraham, N. S. (2015). Literature Survey On Various Region Extraction Methods. *International Journal of Engineering Research and General Science*, *3*(4), 1156–1159.

AFB. (2003). *Enlarging the View: ZoomText and MAGic Screen-Magnification Software, Part 2*. American Foundation for the Blind. (AFB). Retrieved from https://www.afb.org/aw/4/5/14820 on 22/03/2019.

Agarwal, D. (2015). A Comparative Analysis of Edge Detection Techniques Used in Flame Image Processing. *International Journal of Advance Research In Science And Engineering IJARSE, 4*(1), 1335–1343.

Aiswarya. L. K, & Shiji.T. P. (2019). Design and Construction of Electronic Aid for Visually Impaired People. *International Journal of Scientific Development and Research (IJSDR), 4*(4), 110–113.

Al-Shehabi, M. M., Mir, M., Ali, A. M., & Ali, A. M. (2014). An Obstacle Detection and Guidance System for Mobility of Visually Impaired in Unfamiliar Indoor Environments. *International Journal of Computer and Electrical Engineering*, *6*(4), 337–341. Retrieved from https://doi.org/10.7763/IJCEE.2014.V6.849 on 12/07/2021.

Aladr´en, A., L´opez-Nicol´as, G., & Guerrero, L. P. (2016). Navigation Assistance for the Visually Impaired Using RGB-D Sensor with Range Expansion. *IEEE Systems Journal*, *10*(3), 922–932.

Alam, N., Islam, M., Habib, A., & Mredul, M. B. (2018). Staircase Detection Systems for the Visually Impaired People : A Review. *International Journal of Computer Science and Information Security.(IJCSIS)*, *16*(12), 13–18.

Alghamdi, S., Schyndel, R. G. van, & Khalil, I. (2014). Accurate positioning using

long range active RFID technology to assist visually impaired people. *Journal of Network and Computer Applications*, *41*(1), 135–147. Retrieved from https://doi.org/10.1016/j.jnca.2013.10.015 on 12/07/2021.

Ansari, M. A., Kurchaniya, D., & Dixit, M. (2017). A comprehensive analysis of image edge detection techniques. International Journal of Multimedia and Ubiquitous Engineering, 12(11), 1-12.

Aravind P. (2020). CNN vs. RNN vs. ANN – Analyzing 3 Types of Neural Networks in Deep Learning. Retrieved from https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/ on 5/ 10/2021

Aviram, Z. (2018). *OrCam CEO/Co-Founder, Ziv Aviram, on BBC, Discussing OrCam MyEye 2*. BBC News. Retrieved from https://www.orcam.com/en/media/orcam-ceo-and-co-founder-ziv-aviram-on-bbc-discussing-orcam-myeye-2-the-most-advanced-wearable-assistive-device-for-the-blind-and-visually-impaired/ on 14/08/2019.

Azlina A. (2016). The value of green space to people with a late onset visual impairment: A case of people with Age-related Macular Degeneration (AMD) in Scotland, United Kingdom. University of Edinburgh. Retrieved on 17/06/2020.

Babu, S., Suryakeerthi, V., Raju, S. M., Mathew, P., & George, R. M. (2015). *Wearable Device for Blind*. *3*(05), 1–5.

Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9365-9374.

Bai, J., Lian, S., Liu, Z., Wang, K., & Liu, D. (2017). Smart Guiding Glasses for

Visually Impaired People in Indoor Environment. *In IEEE Trans. Consum. Electron.*, *63*(3), 258–266.

Bai, J., Liu, Z., Lin, Y., Li, Y., Lian, S., & Dijun Liu. (2019). Wearable Travel Aid for Environment Perception and Navigation of Visually Impaired People. *Electronics*, *8*(1), 697. Retrieved from https://doi.org/10.3390/electronics8060697 on 23/04/2021

Barati, F., & Delavar, M. R. (2015). Design and development of a mobile sensor based the blind assistance wayfinding system. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, *40*(1W5), 91–96. Retrieved from https://doi.org/10.5194/isprsarchives-XL-1-W5-91-2015 on 16/05/2021.

Bharambe, S., Thakker, R., Patil, H., & Bhurchandi, K. M. (2013). Substitute eyes for Blind using Android Abstract : *Texas Instruments India Educators' Conference*, 38–43. Retrieved from https://doi.org/10.1109/TIIEC.2013.14 on 18/05/2021.

Bhatlawande, S., Ieee, S. M., Mahadevappa, M., & Ieee, S. (2014). Design , Development and Clinical Evaluation of the Electronic Mobility Cane for Vision Rehabilitation. *IEEE Trans. Neural Syst. Rehabil. Eng.*, *22*(6), 1148–1159. Retrieved from https://doi.org/10.1109/TNSRE.2014.2324974 on 19/05/2021

Bhowmick, A. (2017). An insight into assistive technology for the visually impaired and blind people : state-of-the-art and future trends. *Journal on Multimodal User Interfaces*, 2(5). Retrieved from https://doi.org/10.1007/s12193-016-0235-6 on 23/11/2019.

Borenstein, J., & Ulrich, I. (1997, April). The guidecane-a computerized travel aid

for the active guidance of blind pedestrians. In Proceedings of International Conference on Robotics and Automation IEEE, 2, 1283-1288.

Bousbia-salah, M., & Fezari, M. (2007). A Navigation Tool for Blind People. *Innovations and Advanced Techniques in Computer and Information Sciences and Engineering*, 333–337. Retrieved from https://doi.org/10.1007/978-1-4020-6268-1on 20/09/2019

Brock, A. (2013). Interactive maps for visually impaired people: design, usability and spatial cognition (Doctoral dissertation, Université Toulouse 3 Paul Sabatier).

Cardillo, E., Mattia, V. Di, Manfredi, G., Russo, P., Leo, A. De, Caddemi, A., & Cerri, G. (2018). An Electromagnetic Sensor Prototype to Assist Visually Impaired and Blind People in Autonomous Walking. *IEEE Sensors Journal*, *18*,(6), 2568–2576. Retrieved from https://doi.org/10.1109/JSEN.2018.2795046 on 17/10/2020

Cecílio, J., Duarte, K., & Furtado, P. (2015). BlindeDroid : An Information Tracking System for Real-time Guiding of Blind People. *ELSEVIER*, *5(2)*, 113–120. Retrieved from https://doi.org/10.1016/j.procs.2015.05.039 on 18/11/2019

Chauhan, S., Sharma, E., & Doegar, A. (2016). Binarization Techniques for Degraded Document Images - A Review. *International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, 163–166.

Chavre, P. B. (2015). A Survey on Text Localization Method in Natural Scene Image A Survey on Text Localization Method in Natural Scene Image. *International Journal of Computer Applications*, *112*(13), 0975 – 8887.

Cheraghi, S. A., Vinod, N., & Walker, L. (2017). GuideBeacon: Beacon-based indoor wayfinding for the blind, visually impaired,and disoriented. *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 121–130. Retrieved from https://doi.org/10.1109/PERCOM.2017.7917858 on 15/09/2021

Chukwunazo, E. J., & Onengiye, G. M. (2016). Design and implementation of microcontroller based mobility aid for visually impaired people. International journal of science and research (Nagpur), 5(6), 680-686.

Crudden, A. (2002). Employment after Vision Loss: Results of a Collective Case Study. *Journal of Visual Impairment & Blindness (JVIB)*, *96*(9), 615–621.

Dakopoulos, D., & Bourbakis, N. G. (2010). For Blind : A Survey. *IEEE Transactions on Systems, Man, and Cybernetics*, *40*(1), 25–35.

Das, S. (2016). Comparison of Various Edge Detection Technique. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, *9*(2), 143–158. Retrieved from http://dx.doi.org/10.14257/ijsip.2016.9.2.13 Comparison on 15/09/2021

Dey, A. (2016). Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174–1179. www.ijcsit.com Retrieved from 16/09/2021.

Diwakara, M., & Surendran, D. (2018). Real Time Indoor Path Navigation. *International Journal of Pure and Applied Mathematics*, *119*(12), 1339–1346.

Doermann D, Liang J, L. H. (2003). Progress in camera-based document image analysis. *Seventh International Conference on Document Analysis and Recognition, 2003 Proceedings*, 606–616.

Donnell, W. O. (2014). An Analysis of Employment Barriers Facing Blind People.

Retrieved from https://scholarworks.umb.edu/mspa_capstone/23/ on 28/04/2019.

Duquette, J., & Baril, F. (2013). Factors influencing work participation for people with a visual impairment. Retrieved from https://extranet.inlb.qc.ca/wp-content/uploads/2015/01/Factors-influencing-work-participation-in-persons-with-VI.pdf on 23/07/2019.

Elgendy, M., Sik-lanyi, C., & Kelemen, A. (2019). Making Shopping Easy for People with Visual Impairment Using Mobile Assistive Technologies. *Journal of Applied Sciences*, *9*(1), 1–15. Retrieved from https://doi.org/10.3390/app9061061 on 24/09/2021.

Elmannai, W. M., & Elleithy, K. M. (2018). A Highly Accurate and Reliable Data Fusion Framework for Guiding the Visually Impaired. *IEEE Access*, *6*, 33029–33054. Retrieved from https://doi.org/10.1109/ACCESS.2018.2817164 on 20/11/2020.

Emily, E. O., Mohtar, A. A., Diment, L. E., & Reynolds, K. J. (2014). A Detachable Electronic Device for Use With a Long White Cane to Assist With Mobility.*Assistive Technology*, *26*(4), 219–226. doi:10.1080/10400435.2014.926468.

Ercoli, I., Marchionni, P., & Scalise, L. (2013, September). A wearable multipoint ultrasonic travel aids for visually impaired. *In Journal of physics: conference series* (Vol. 459, No. 1, p. 012063). IOP Publishing. Retrieved from https://doi.org/10.1088/1742-6596/459/1/012063 on 20/11/2020.

Eunjeong, K., & Eun, Y. K. (2017). A Vision-Based Wayfinding System for Visually Impaired People Using Situation Awareness and Activity-Based Instructions. *Sensors*, *17*, 1882. Retrieved from

https://doi.org/10.3390/s17081882 on 20/09/2021.

Fawaz, A., Zaman, N., Qazi, M., Faouda, Y., & Ahmad, M. (2013). Proposing Smart System For Blind. *Journal of Applied Sciences*, *13*(7), 1112–1117.

Fei, L., Wang, K., Lin, S., Yang, K., Cheng, R., & Chen, H. (2018). Scene text detection and recognition system for visually impaired people in real world. *SPIEDigitalLibrary.Org/Conference-Proceedings-of-Spie*, 10. Retrieved from https://doi.org/10.1117/12.2325523 on 21/02/2021

Gao, J., & Yang, J. (2001). An Adaptive Algorithm for Text Detection from Natural Scenes. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.*, *2*, II–II.

Garcia-macias, J. A., Ramos, A. G., Saul, E., & Hernandez, P. (2019). Uasisi : a modular and daptable wearable system to assist the visually impaired. Procedia Computer Science,151, 425–430. Retrieved from https://doi.org/10.1016/j.procs.2019.04.058 on 25/04/2022

Gawari, H., & Bakuli, M. (2014). Voice and GPS Based Navigation System for Visually Impaired. *International Journal of Engineering Trends and Technology*, *11*(6), 296–299. Retrieved from https://doi.org/10.14445/22315381/ijett-v11p256 on 23/06/2019.

Guerrero, L. A., Vasquez, F., & Ochoa, S. F. (2012). An indoor navigation system for the visually impaired. *Sensors (Switzerland)*, *12*(6), 8236–8258. Retrieved from https://doi.org/10.3390/s120608236 on 13/10/2021

Gupta, N., & Banga, V. K. (2012). Localization of Text in Complex Images Using Haar Wavelet Transform. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *1*(6), 111–115.

Habeeb, A. (2017). Artificial intelligence Ahmed Habeeb University of Mansoura.

Retrieved from https://doi.org/10.13140/RG.2.2.25350.88645 on 28/13/2019

Hanley, D. (2014). Conceptualising disability in the workplace : Contextualising the responses of managers and employees [Central Lancashire]. retrieved from http://clok.uclan.ac.uk/10716/2/Hanley David Final e-Thesis %28Master Copy%29.pdf on 14/03/2021.

Hinde, A., & Elfkihi, S. (2010). Features Extraction for Text Detection and Localization . *In the 5th IEEE International Symposium On I/V Communications and Mobile Network*, 1–4. doi:10.1109/ISVC.2010.5656284.

Hussin, A. (2013). Experiences of students with visual impairments in adoption of digital talking textbooks: An interpretative phenomenological analysis (Doctoral dissertation, Colorado State University).

Islam, M., & Sadi, M. S. (2018). Path Hole Detection to Assist the Visually Impaired People in Navigation. *4th International Conference on Electrical Engineering and Information & Communication Technology (ICEEiCT)*, 268–273.

Islam, M., Sadi, M. S., Member, S., & Zamli, K. Z. (2019). Developing Walking Assistants for Visually Impaired People : A Review. *IEEE Sensors Journal*, *19*(8), 2814–2828. Retrieved from https://doi.org/10.1109/JSEN.2018.2890423 on 12/03/2021

Joseph, S. L., Xiao, J., & Zhang, X. (2015). Being Aware of the World : Toward Using Social Media to Support the Blind. *IEEE Transactions on Human-Machine Systems*, *45*(3), 1–7. Retrieved from https://doi.org/10.1109/THMS.2014.2382582 on 17/03/2021

Jung, K. (2001). Neural network-based text location in color images. *Pattern Recognition Letters, - Elsevier*, *22*(14), 1503–1515.

Kamal, M. M., Bayazid, A. I., Sadi, M. S., Islam, M. M., & Hasan, N. (2017, December). Towards developing walking assistants for the visually impaired people. *In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)* (pp. 238-241). IEEE.

Kambala, M. S., Haritha, C. P., Kasim, B., G S Roja, P., & G Sankara, R. (2020). Optical Character Recognition using CRNN. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *9*(8), 115–120. Retrieved from https://doi.org/10.35940/ijitee.H6264.069820 on 12/04/2021

Kammoun, S., Parseihian, G., Gutierrez, O., Brilhault, A., Serpa, A., Raynal, M., Oriola, B., Macé, M. J., Auvray, M., Denis, M., Thorpe, S. J., Truillet, P., Katz, B. F. G., & Jouffrais, C. (2012). Navigation and space perception assistance for the visually impaired. *IRBM*, *33*(2), 182–189. Retrieved from https://doi.org/10.1016/j.irbm.2012.01.009 on 29/04/2021.

Kantipudi, M. V. V., Kumar, S., & Kumar Jha, A. (2021). Scene text recognition based on bidirectional LSTM and deep neural network. *Computational Intelligence and Neuroscience*, *2021*.

Karande, P. ;, & Jogdand, S. (2018). Portable Camera Based Assistive Text Reading from Hand Held Objects for Blind Person. *International Journal of Advanced Research in Computer and Communication Engineering*, *5*(6), 625–628. Retrieved from https://doi.org/10.17148/IJARCCE.2016.56137. on 26/05/2021.

Katzschmann, R. K., Araki, B., & Rus, D. (2018). Safe local navigation for visually impaired users with a time-of-flight and haptic feedback device. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *26*(3), 583–593.

Kevadia, R., Tushar, R., Ayushi, Ra., Mayank, K., & D, P. (2018). Wearable Technology for Visually Impaired. *International Journal of Innovative Research in Engineering &Management(IJIREM)*, *5*(2), 91–93.

Kevin Doughty. (2016). *Smart Assistive Technology for People with Vision Impairment*. Retrieved from https://www.researchgate.net/publication/306402826_Smart_Assistive_Technology_for_People_with_Vision_Impairment on 23/08/2020.

Khaliluzzaman, M., Deb, K., & Jo, K. (2016). Stairways detection and distance estimation approach based on three connected point and triangular similarity. *9th International Conference on Human System Interactions (HSI)*, 330–336.

Khan, A., Usmani, M. N., Rahman, N., & Prasad, D. (2019). Pre-Processing Images of Public Signage for OCR Conversion. *Journal of Signal and Information Processing*, *10*(1), 1–11. https://doi.org/10.4236/jsip.2019.101001

Kiuru, T., Metso, M., Utriainen, M., Metsävainio, K., Jauhonen, H., Rajala, R., Savenius, R., Ström, M., Jylhä, T., Juntunen, R., & Sylberg, J. (2018). Assistive device for orientation and mobility of the visually impaired based on millimeter wave radar technology — Clinical investigation results. *Cogent Engineering*, *13*(1), 1–12. https://doi.org/10.1080/23311916.2018.1450322

Kulyukin, V., & Kutiyanawala, A. (2010). Accessible Shopping Systems for Blind and Visually Impaired Individuals : Design Requirements and the State of the Art Accessible Shopping Systems for Blind and Visually Impaired Individuals : Design Requirements and the State of the Art. *Open Rehabilitation Journal*, *3*(1), 158–168. https://doi.org/10.2174/1874943701003010158

Kumar, G. A. E. S., & Munindhar, M. (2019). Intelligent Assistive System for

Visually Disabled Persons. *International Journal of Recent Technology and Engineering(IJRTE)*, *8*(4), 1436–1440. https://doi.org/10.35940/ijrte.D7404.118419

Kumar, M., & Saxena, R. (2013). A Lgorithm T echnique On Various Edge Detection : A S Urvey. Signal & Image Processing *: An International Journal (SIPIJ) Vol.4*, *4*(3), 65–75.

Lazar, J., Allen, A., Kleinman, J., & Malarkey, C. (2007). What Frustrates Screen Reader Users on the Web : A Study of 100 Blind Users. *International Journal Of Human–Computer Interaction*, *22*(3), 247–269.

Lee, C. S., Lee, J. I., & Seo, H. E. (2021). Deep Learning Based Mobile Assistive Device for Visually Impaired People. *2021 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, 1–3. Retrieved from https://doi.org/10.1109/ICCE-Asia53811.2021.9641925 on 7/10/2022

Legge, G. E., Beckmann, P. J., Tjan, B. S., Havey, G., Kramer, K., Rolkosky, D., Gage, R., Chen, M., Puchakayala, S., & Rangarajan, A. (2013). Indoor Navigation by People with Visual Impairment Using a Digital Sign System. *Journal of PLOSONE*, *8*(10), 1–15. https://doi.org/10.1371/journal.pone.0076783

Li, J., Pan, J., & Chu, S. (2008). Kernel class-wise locality preserving projection. *Information Sciences*, *178*, 1-51825–51835. Retrieved from https://doi.org/10.1016/j.ins.2007.12.001 on 20/09/2021.

Lin, B. S., Lee, C. C., & Chiang, P. Y. (2017). Simple smartphone-based guiding system for visually impaired people. *Sensors (Switzerland)*, *17*(6), 1371. https://doi.org/10.3390/s17061371

López-de-ipiña, D., Lorido, T., & López, U. (2011). BlindShopping : Enabling

Accessible Shopping for Visually Impaired People through Mobile Technologies BlindShopping. *Research Gate*, 266–270. Retrieved from https://doi.org/10.1007/978-3-642-21535-3 on 14/02/2020.

Mahmood, H. (2020). Text-detection and-recognition from natural images (Doctoral dissertation, Loughborough University).

Mahmood, H. F., & Edirisinghe, E. (2017). Text Localization in Natural Images Through Effective Re - Identification of the MSER. *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, 1–9.

Mancini, A., Zingaretti, E., & Zingaretti, P. (2018). Mechatronic system to help visually impaired users during walking and running. *IEEE Trans. Intell. Transp. Syst.*, *19*(2), 649–660.

Martin, E. (2016). *Technology Based Aid for the Visually Impaired* (Doctoral dissertation, Thesis, The University of Dublin, Trinity College Dublin). Retrieved from https://scss. tcd. ie/publications/theses/diss/2016/TCD-SCSS-DISSERTATION-2016-007. pdf).

Martinez-sala, A. S., Losilla, F., Sánchez-aarnoutse, J. C., & García-haro, J. (2015). Design , Implementation and Evaluation of an Indoor Navigation System for Visually Impaired People. *Sensors (Basel)*, *15*(12), 32168–32187. https://doi.org/10.3390/s151229912

Mehta, P., Kant, P., Shah, P., & Roy, A. K. (2011, June). VI-Navi: a novel indoor navigation system for visually impaired people. *In Proceedings of the 12th International Conference on Computer Systems and Technologies* 365-371.

Mirmehdi, M., Palmer, P. L., & Kittler, J. (1997, June). Towards optimal zoom for automatic target recognition. *In Proceedings of the Scandinavian Conference*

on Image Analysis (Vol. 1, pp. 447-454). Proceedings Published by Various Publishers.

Mishra, A., & Vedhapriyavadhana, R. (2022). Medicine assistance application for visually impaired people. *Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT)*, 850–855. https://doi.org/10.1109/ICICICT54557.2022.9917621

Mohamad, M. H., Rana, S., & Sayemul Islam. (2013). Smart walking stick-an electronic approach to assist visually disabled persons. *International Journal of Scientific & Engineering Research*, *4*(10), 111–114. https://doi.org/10.1088/1748-6041/6/5/055002

Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). *Machine learning: algorithms and applications*. Crc Press.

Mohanty, S., Karunan, M., Sayyad, I., Khursade, S., & Gite, P. B. B. (2017). Smart Shoe for Visually Impaired. *International Journal of Advan Ced Research in Computing and Communication Engineering (IJARCCE)*, *6*(11), 244–246. https://doi.org/10.17148/IJARCCE.2017.61138

Mohler, C. E. (2012). *The Process Of Obtaining And Retaining Employment Among The Vision-restricted* [University of Western Ontario London, Ontario, Canada ©]. Retrieved from http://ir.lib.uwo.ca/digitizedtheses on 16/05/2019

Mollah, A. F., Majumder, N., Basu, S., & Nasipuri, M. (2011). Design of an Optical Character Recognition System for Camera-based Handheld Design of an Optical Character Recognition System for Camera- based Handheld Devices. *IJCSI International Journal of Computer Science Issues*, *8*(4), 1694–0814.

Morad, A. H. (2010). GPS Talking For Blind People. J*ournal of Emerging*

Technologies in Web Intelligence, *2*(3) ,239-243. https://doi.org/10.4304/jetwi.2.3.239-243

Muhammad, A. (2012). Smartphone based guidance system for visually impaired person Smartphone based Guidance System for Visually Impaired Person. *Research Gate*, 1–7. Retrieved from https://doi.org/10.1109/IPTA.2012.6469553 on 03/10/2019

Muthusenthil B, Joshuwa J, Kishore S, Narendiran K. (2018). Smart Assistance for Blind People using Raspberry Pi. *International Journal of Advance Research, Ideas and Innovations in Technology*, *4*(2), 884–891.

Nadhir, M., Wahab, A., Sufril, A., Mohamed, A., & Abdull, A. S. (2021). *Text Reader for Visually Impaired Person*. Retrieved from https://doi.org/10.1088/1742-6596/1755/1/012055 on 09/03/2022

Donaldson, N. (2017). *Visually impaired individuals' perspectives on obtaining and maintaining employment (*Doctoral dissertation, Walden University).

Nair, V., Tsangouri, C., Xiao, B., Olmschenk, G., Seiple, W. H., & Zhu, Z. (2018). A hybrid indoor positioning system for blind and visually impaired using Bluetooth and Google tango. *Journal on technology and persons with disabilities*, 6, 63-82.

Nixon, M. S., & Aguado, A. S. (2020). *Feature Extraction and Image Processing for Computer Vision Fourth Edition*. Elsevier Ltd. Retrieved from https/:www.elsevier.com/books and journals, on 11/07/2021

Noman, A. T., Chowdhury, M., & Rashid, H. (2017). Design and Implementation of Microcontroller Based Assistive Robot for Person with Blind Autism and Visual Impairment. *International Conference of Computer and Information Technology (ICCIT)*,5–10. Retrieved from

https://doi.org/10.1109/ICCITECHN.2017.8281806 on 23/06/2020.

Patel, S., & Makwana, R. (2020). Connectionist Temporal Classification Model for Dynamic Hand Gesture Recognition using RGB and Optical flow Data. *The International Arab Journal of Information Technology*, *17*(4), 497–506.

Poornima, B., Ramadevi, Y., & Sridevi, T. (2011). Threshold based edge detection algorithm. *International Journal of engineering and technology*, 3(4), 400-403.

Pruettikomon, S., & Louhapensang, C. (2018). A study and development of workplace facilities and working environment to increase the work efficiency of persons with disabilities: A Case Study Of Major Retail And Wholesale Companies in Bangkok. *The Scientific World Journal*, 2018(2):1-12. DOI: 10.1155/2018/3142010 2018.

Pruthvi, S., Nihal, P. S., Menon, R. R., Kumar, S. S., & Tiwari, S. (2019). Smart Blind Stick using Artificial Intelligence. *International Journal of Engineering and Advanced Technology (IJEAT)*, *8*(5), 19–22.

Ramadhan, A. J. (2018). Wearable smart system for visually impaired people. *Sensors (Switzerland)*, *18*(3), 843. Retrieved from https://doi.org/10.3390/s18030843 on 23/03/2020 on 24/03/2021

Ravi Kuber, Amanda Hastings, M. T. (2012). Determining the Accessibility Of Mobile Screen Readers For Blind Users. *Proceedings of IASTED Conference on Human-Compuer Interface Behaviour, USA*, 182–189.

Raymond B. H., & Alcides A. S. (2010). Basic Human Computer Interface for the Blind. *Eighth LACCEI Latin American and Caribbean Conference for Engineering and Technology (LACCEI'2010) "Innovation and Development for the Americas"*, 1–10.

Romić, K., Irena, G., & Tomislav, G. (2015). Technology assisting the blind — Video processing based staircase detection. *In Proc. 57th International Symposium ELMAR (ELMAR)*, 221–224 Retrieved from https://doi.org/10.1109/ELMAR.2015.7334533 on 26/09/2021.

Sahoo, N., Lin, H., & Chang, Y. (2019). Design and Implementation of a Walking Stick Aid. *Sensor*, *19*(1), 130. https://doi.org/10.3390/s19010130

Salam, M. Al, & Korial, A. E. (2016). Indoor Navigation for Visually Impaired / Blind People Using Smart Cane and Mobile Phone : Experimental Work Indoor. *Journal of Information Engineering and Applications*, *6*(5), 2224–5782.

Santiago R., & Araujo, A. (2019). Navigation systems for the blind and visually impaired: Past work, challenges, and open problems. *Sensors*, 19(15), 3404.

Sarojini, L., Anburaj, I., Aravind, R., Karthikeyan, M., & Gayathri, K. (2017). Smart electronic gadget for visually impaired people. *Indian J.Sci. Res.*, *14*(1), 348–352.

Sato, D., & Kitani, K. (2017). NavCog3 : An Evaluation of a Smartphone-Based Blind Indoor Navigation Assistant with Semantic Features in a Large-Scale Environment. *ASSETS*, *17*, 270–279.

Sawaki, M., Murase, H., & Hagita, N. (2000). Automatic Acquisition of Context-based Images Templates for Degraded Character Recognition in Scene Images. *In Proceedings 15th International Conference on Pattern Recognition. ICPR-*, *4*, 15–18.

Shah H., (2015). A Prototype of Assistive Text Label Reading System for Blind – An Overview. *IJSRD - International Journal for Scientific Research & Development*, *3*(01), 1040–1042.

Shikha, C., Singhal, V., & Sonu, M. (2020). Ancient Text Character Recognition using Deep Learning. *International Journal of Engineering Research and Technology*, *13*(9), 2177–2184. Retrieved from https://dx.doi.org/10.37624/IJERT/13.9.2020.2177-2184 on 17/06/2021

Shrivakshan, G. T. (2012). A Comparison of various Edge Detection Techniques used in Image Processing. *IJCSI International Journal of Computer Science Issues*, *9*(5), 269–276.

Smith, P., Reid, D. B., Environment, C., Palo, L., Alto, P., & Smith, P. L. (1979). A Tlreshold Selection Method from Gray-Level Histograms. *EEE Transactions On Systrems, Man, And Cybernetics*, *9*(1), 62–66.

Sohl-dickstein, J., Teng, S., Gaub, B. M., Rodgers, C. C., Li, C., Deweese, M. R., & Harper, N. S. (2015). A device for human ultrasonic echolocation. *IEEE Trans. Biomed. Eng.*, *62*(6), 1526–1534.

Sotomayor, N., Danilo, G., Garcia, C., Camacho, O., & Sarzosa, M. (2019). An approach for Helping the Mobility of People with Visual Impairment : Design and Implementation. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, *23*(10), 194–205.

Su, B., Member, S. L., Lim, C., & Senior, T. (2012). A Robust Document Image Binarization Technique for Degraded Document Images. *IEEE Transactions on Image Processing*, *22*(4), 1408–1417.

Suchismita S.(2021). Decision Boundary for Classifiers: An Introduction. Retrieved from https://medium.com/analytics-vidhya/https://medium.com/analytics-vidhya/decision-boundary-for-classifiers-an-introduction-cc67c6d3da0edecision-boundary-for-classifiers-an-introduction-cc67c6d3da0%20on%2019/03/2022. on 19/03/2022.

Tapu, R., Mocanu, B., Bursuc, A., & Zaharia, T. (2013). A Smartphone-based Obstacle Detection and Classification System for Assisting Visually Impaired People. *Proceedings of the IEEE International Conference on Computer Vision*, 444–451. Retrieved from https://doi.org/10.1109/ICCVW.2013.65 on 23/09/2021.

Tejaswani, G., Afroz, B., & Sunitha, S. (2018). A Text Recognizing Device For Visually Impaired. *International Journal of Engineering and Computer Science.(IJECS)*, *7*(3), 23697–23700. https://doi.org/10.18535/ijecs/v7i3.07

TheAILearner. (2022). *Convolutional Recurrent Neural Network for Text Recognition*. Retrieved from https://Keras/blob/master/CRNN Model.ipynb on 11/09/2021.

Ti, Y. (2022). Scene Text Detection and Recognition by CRAFT and a Four-Stage Network. 0–6. Retrieved from https://www.researchgate.net/publication/357577698_Scene_Text_Detection_and_Recognition_by_CRAFT_and_a_Four-Stage_Network, on 08/09/2022.

Trémeau, A., Godau, C., Karaoglu, S., & Muselet, D. (2011). Detecting Text in Natural Scenes Based on a Reduction of Photometric Effects : Problem of Color Invariance. *In International Workshop on Computational Color Imaging*, *April*, 214–229. https://doi.org/10.1007/978-3-642-20404-3

Trivedi, A., Pant, N., Shah, P., Sonik, S., & Agrawal, S. (2018). Speech to text and text to speech recognition systems-Areview. *IOSR Journal of Computer Engineering (IOSR-JCE)*, *20*(2), 39. https://doi.org/10.9790/0661-2002013643

Uchida, S. (2020). Text Localization and Recognition in Images and Video. Retrieved from https://doi.org/10.1007/978-0-85729-859-1 on 27/10/2021.

Ugwu, C., Wiliams, E. E., & Nwachukwu, E. O. (2012). *Introduction to Artificial Intelligence and Expert Systems* (1st ed.). MunaGenesis Concepts Nig. Ltd.

Ullah, S., Ahmed, S., Ahmad, W., & Jan, L. (2018). Sensor Based Smart Stick for Blind Person Navigation. *International Journal of Scientific & Engineering Research(IJSER)*, *9*(12), 106–109.

Uma, K., Dagade, R., & Shiravale, S.; (2018). Maximally Stable Extremal Region Approach for Accurate Text Detection in Natural Scene Images. *International Journal of Scientific Development and Research. IJSDR*, *1*(11), 161–168.

Valsan, K. S. (2017). Smart Braille Recognition System. *International Journal for Research in Applied Science and Engineering Technology*, *V*(XI), 2452–2460. https://doi.org/10.22214/ijraset.2017.11343

Vaz, R., Fernandes, P. O., Cecília, A., & Veiga, R. (2018). Designing an Interactive Exhibitor for Assisting Blind and Visually Impaired Visitors in Tactile Exploration of Original Museum Pieces Impaired Visitors in Tac. *Procedia Computer Science*, *138*, 561–570. Retrieved from https://doi.org/10.1016/j.procs.2018.10.076 on 26/07/2021

Vel, R., & Rodrigo, P. (2018). An Outdoor Navigation System for Blind Pedestrians using GPS and Tactile-Foot Feedback. *Applied Sciences*, *8*, 578. Retrieved from https://doi.org/10.3390/app8040578 on 25/09/2020

Vijeesh, V., Kaliappan, E., Ponkarthika, B., & Vignesh, G. (2019). *Design of Wearable shoe for blind using LIFI Technology*. *8*(6), 1398–1401. https://doi.org/10.35940/ijeat.F1248.0986S319

Wafa, E., & Elleithy, K. (2017). Sensor-Based Assistive Devices for Visually-Impaired People: Current Status, Challenges, and Future Directions. *Sensors*, *17*, 565. https://doi.org/10.3390/s17030565

Wang, S., & Tian, Y. (2012). Detecting stairs and pedestrian crosswalks for the blind by RGBD camera Detecting Stairs and Pedestrian Crosswalks for the

Blind by RGBD Camera. *In Proc. IEEE International Conference on Bioinformatics and Biomedicine Workshops*, 732–739. Retreived from https://doi.org/10.1109/BIBMW.2012.6470227 on 20/10/2020.

Wang, X., Song, Y., & Zhang, Y. (2013). Natural Scene Text Detection with Multi-channel Connected Component Segmentation. *Institute of Artificial Intelligence and Robotics Xi'an*. Retrieved from https://doi.org/10.1109/ICDAR.2013.278 on 27/10/2020

Yadav, A., Sharma, B., & Pal, A. (2018). Smart Guiding System for Blind : Obstacle Detection and Real-time Assistance via GPS. *International Research Journal of Engineering and Technology(IRJET), 5*(3), 500–502.

Yadav, N., & Kumar, V. (2016). A Review of various Text Localization Techniques. Retrieved from https://hal.inria.fr/hal-01326642 on 23/08/2019.

Yang, K., Wang, K., Hu, W., & Bai, J. (2016). Expanding the Detection of Traversable Area with RealSense for the Visually Impaired based in stereo. *Sensors*, *16*(11), 1954. https://doi.org/10.3390/s16111954

Ye, Q., & Doermann, D. (2015). Text Detection and Recognition in Imagery : A Survey. Retrieved from https://doi.org/10.1109/TPAMI.2014.2366765 on 20/14/2021.

Yeonju O., Kao, W. L., & Min, B. C. (2017, July). Indoor navigation aid system using no positioning technique for visually impaired people. *In International Conference on Human-Computer Interaction* (pp. 390-397). Springer, Cham.

Yi, C., & Tian, Y. (2012). Assistive text reading from complex background for blind persons. *In International Workshop on Camera-Based Document Analysis and Recognition*, 15–18.

Zandifar A., Duraiswami R.,& Chahine A. (2002). A video based interface to

textual information for the visually impaired. *Proceedings Fourth IEEE International Conference on Multimodal Interfaces*, 25–30.

Zhigang, Z., Princeton, N., Tony, R., Edgardo, M., Frank, P., & Wai, K. (2014). Wearable Navigationassistance for the Vision-Impared. *United States Patent Application Publication*, *1*(19), 1–11.

Zhou, D., Yang, Y., & Yan, H. (2016). A Smart 'Virtual eye' Mobile System for the Visually Impaired. *IEEE Potentials*, *35*(6), 13–20.

Difference between ANN, CNN and RNN. Retrieved from https://www.geeksforgeeks.org/difference-between-ann-cnn-and-rnn/. on 5/10/2021.

Blindness and Vision Impairment: Retrieved from https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment on 13/10/2022

Synth90k synthetic text dataset by Visual Geometry Group. Retrieved from https://www.robots.ox.ac.uk/~vgg/data/text/ on 3/10/ 2021

APPENDIX I

SOURCE CODE LISTING

Data Preparation

#1. Fetch the whole mjsynth dataset and unzip 15GB of the total 21GB:

```
!curl -LO https://thor.robots.ox.ac.uk/~vgg/data/text/mjsynth.tar.gz
!tar -xf mjsynth.tar.gz
```

#2. Create three variables to hold for training, validation and test datasets and set header to None:

```
train = pd.read_csv(dir_path+'/annotation_train.txt', header = None)
validation = pd.read_csv(dir_path+'/annotation_val.txt', header = None)
test = pd.read_csv(dir_path+'/annotation_test.txt', header = None)
```

#3. Check for the length of training, validation and test variables:

```
Original Train Dataset Size : 7224612
Original Validation Dataset Size : 802734
Original Test Dataset Size : 891927
```

#4. Output format of dataframe in train variable:

	0
0	./2425/1/115_Lube_45484.jpg 45484
1	./2425/1/114_Spencerian_73323.jpg 73323
2	./2425/1/113_accommodatingly_613.jpg 613
3	./2425/1/112_CARPENTER_11682.jpg 11682
4	./2425/1/111_REGURGITATING_64100.jpg 64100

#5. Create actual directories for the Dataset and to hold train, validation and test datasets:

```
# Creating the Directories
! mkdir Data
! mkdir Data/train
! mkdir Data/validation
! mkdir Data/test
```

#6. Move image paths into their designated folders in the Data folder:

```
for path in tqdm(train['path'].values):
    if os.path.isfile(path)==False:
      pass
    else:
      os.rename(path, 'Data/train/'+path.split('/')[-1])

for path in tqdm(validation['path'].values):
  if os.path.isfile(path)==False:
      pass
  else:
      os.rename(path, 'Data/validation/'+path.split('/')[-1])

for path in tqdm(test['path'].values):
  if os.path.isfile(path)==False:
      pass
  else:
    os.rename(path, 'Data/test/'+path.split('/')[-1])
```

#7. Extract labels embedded in path and create a dataframe for both paths and labels:

	path	Label	
0	Data/train/57_babes_5257.jpg	BABES
1	Data/train/358_twinkies_81407.jpg	TWINKIES
2	Data/train/222_BIOCHEMICALS_7556.jpg	BIOCHEMICALS
3	Data/train/387_flecked_29483.jpg	FLECKED
4	Data/train/403_CHATLINES_12870.jpg	CHATLINES

Data Cleaning.

```
train = pd.read_csv('Data/train.csv', na_filter=False)
validation = pd.read_csv('Data/validation.csv', na_filter=False)
test = pd.read_csv('Data/test.csv', na_filter=False
```

check for the number of Numerical and String data in our Label:

```
train_num_labels = []
train_non_num_labels = []

for idx,l in enumerate(tqdm(train.label)):
    if(l.isnumeric()):
        train_num_labels.append(idx)
```

```
    else:
        train_non_num_labels.append(idx)

validation_num_labels = []
validation_non_num_labels = []

for idx,l in enumerate(tqdm(validation.label)):
    if(l.isnumeric()):
        validation_num_labels.append(idx)
    else:
        validation_non_num_labels.append(idx)

test_num_labels = []
test_non_num_labels = []

for idx,l in enumerate(tqdm(test.label)):
    if(l.isnumeric()):
        test_num_labels.append(idx)
    else:
        test_non_num_labels.append(idx)
```

Appendix 11

Preparing Data for Our Recognition Neural Net

#1. Set Parameters for our images

```
batch_size = 16
img_height = 50
img_width = 200
max_label_length = 25
alphabets = set(string.ascii_letters+string.digits)
downsample_factor = 4
epochs = 100
early_stopping_patience = 10
early_stopping = keras.callbacks.EarlyStopping(
    monitor="val_loss", patience=early_stopping_patience, restore_best_weights=T
rue
)
```

#2. Fetch image from Data file:

```
train_df = pd.read_csv('Data/train.csv', na_filter=False)
validation_df = pd.read_csv('Data/validation.csv', na_filter=False)
```

```
test_df = pd.read_csv('Data/test.csv', na_filter=False)

#3. Create image and label variable for train, validation and test datasets:

#images
x_train = train_df.copy()
x_val = validation_df.copy()
x_test = test_df.copy()

#labels
y_train = x_train.pop('label')
y_val = x_val.pop('label')
y_test = x_test.pop('label')

#4. Create character encoder-decoder function:

# Mapping characters to integers
char_to_num = layers.StringLookup(
    vocabulary=list(alphabets), mask_token=None
)

# Mapping integers back to original characters
num_to_char = layers.StringLookup(
    vocabulary=char_to_num.get_vocabulary(), mask_token=None, invert=True
)

def encoder_decoder(img_path, label):
    # 1. Read image
    img = tf.io.read_file(img_path)
    # 2. Decode and convert to grayscale
    img = tf.io.decode_png(img, channels=1)
    # 3. Convert to float32 in [0, 1] range
    img = tf.image.convert_image_dtype(img, tf.float32)
    # 4. Resize to the desired size
    img = tf.image.resize(img, [img_height, img_width])
    # 5. Transpose the image because we want the time
    # dimension to correspond to the width of the image.
    img = tf.transpose(img, perm=[1, 0, 2])
    # 6. Map the characters in label to numbers
    label = char_to_num(tf.strings.unicode_split(label, input_encoding="UTF-8"))
```

```
    # 7. Return a dict as our model is expecting two inputs
    return {"image": img, "label": label}
```

#5. Preprocess dataset for neural net by adding encoder-decoder, batch_size and buffer prefetch:

```
train_dataset = tf.data.Dataset.from_tensor_slices((x_train, y_train))

train_dataset = (
    train_dataset.map( encode_single_sample, num_parallel_calls=tf.data.AUTOTU
NE)
    .batch(batch_size)
    .prefetch(buffer_size=tf.data.AUTOTUNE)
)

validation_dataset = tf.data.Dataset.from_tensor_slices((x_val, y_val))

validation_dataset = (
    validation_dataset.map(
        encode_single_sample, num_parallel_calls=tf.data.AUTOTUNE
    )
    .batch(batch_size)
    .prefetch(buffer_size=tf.data.AUTOTUNE)

)
```

Appendix III

Text Detection and Recognition

```
from numpy.matrixlib.defmatrix import matrix
import cv2
import numpy as np
import docdetect
import time
import re
import pyttsx3
#from skimage.filters import threshold_local
import os
import qdarkstyle
#import keyboard
```

```
# PyQt5 libraries
import sys
from PyQt5.QtWidgets import  QMainWindow, QWidget, QLabel, QApplication,
QMessageBox
from PyQt5.QtCore import QThread, Qt, pyqtSignal, pyqtSlot
from PyQt5.QtGui import QImage, QPixmap
from PyQt5 import QtCore, QtGui, QtWidgets
from PyQt5.QtGui import QMovie
from PyQt5.QtCore import QSize


############### functions #################

# text to speech function:
def speech_function(text):
            engine = pyttsx3.init()
            engine.getProperty('rate')
            voices = engine.getProperty('voices')
            engine.setProperty('voice', voices[1].id)
            engine.setProperty('rate', 118)
            engine.say(text)
            engine.runAndWait()

# speech recognition function:
def takeCommand():

    r = sr.Recognizer()

    with sr.Microphone() as source:

        print("Listening...")
        r.pause_threshold = 1
        audio = r.listen(source)

    try:
        print("Recognizing...")
        query = r.recognize_google(audio, language ='en-in')
        print(f"User said: {query}\n")

    except Exception as e:
        print(e)
        print("Unable to Recognize your voice.")
        return "None"
```

```
    return query

########################## image height and width #################

imgWidth = 640
imgHeight = 480

############################


# Camera read
webcam = True

# reorder points:
def reorder(myPoints):

    myPoints = myPoints.reshape((4, 2))
    myPointsNew = np.zeros((4, 1, 2), dtype=np.int32)
    add = myPoints.sum(1)

    myPointsNew[0] = myPoints[np.argmin(add)]
    myPointsNew[3] =myPoints[np.argmax(add)]
    diff = np.diff(myPoints, axis=1)
    myPointsNew[1] = myPoints[np.argmin(diff)]
    myPointsNew[2] = myPoints[np.argmax(diff)]

    return myPointsNew


# CLASSES:

# speech recognition class:


# vision class:
class ThreadVision(QThread):

    changePixmap = pyqtSignal(QImage)

    def run(self):

        self.ActiveThread = True

        if self.ActiveThread:
```

```
            cap = cv2.VideoCapture(0)
        else:
            print("[INFO] camera not found")

        while True:
            success, img = cap.read()
            img = cv2.normalize(img, img, 0, 255, cv2.NORM_MINMAX)
            cv2.resize(img, (imgWidth, imgHeight))
            imgContour = img.copy()

            imgThreshold = docdetect.pre_processing(img)

            max_cnt = np.array([])
            max_area = 0
            contours, hierarchy = cv2.findContours(imgThreshold,
cv2.RETR_EXTERNAL, cv2.CHAIN_APPROX_NONE)
            for cnt in contours:
                area = cv2.contourArea(cnt)
                if area > 10000:
                    peri = cv2.arcLength(cnt, True)
                    approx = cv2.approxPolyDP(cnt, 0.03 * peri, True)
                    if area > max_area and len(approx) == 4:
                        max_cnt = approx
                        max_area = area
                        cv2.drawContours(imgContour, max_cnt, -1, (255, 0, 0),
20)


            #max_cnt = get_contours(imgThreshold)

            rects = docdetect.process(imgContour)
            img_track = docdetect.draw(rects, imgContour)

            ####################################

            if len(max_cnt) == 4:

                max_cnt = reorder(max_cnt)
                hull = cv2.convexHull(max_cnt)
                perimeter = cv2.arcLength(hull, True)
                corners = cv2.approxPolyDP(hull, 0.03*perimeter, True)

                if corners.size != 8:
                    continue
                else:
```

```
                    factor = img.shape[0]/imgHeight
                    cornerslist =
(corners.reshape(4,2)*factor).astype(int).tolist()

                    tl = min(cornerslist, key = lambda x: x[0]+x[1])
                    tr = max(cornerslist, key = lambda x: x[0]-x[1])
                    br = max(cornerslist, key = lambda x: x[0]+x[1])
                    bl = min(cornerslist, key = lambda x: x[0]-x[1])

                    cornerslist = np.array([tl, tr, br, bl],
dtype='float32').reshape(4,2)
                    print(cornerslist)

                    widthTop = np.sqrt((tl[0] - tr[0])**2 + (tl[1] - tr[1])**2)
                    widthBottom = np.sqrt((bl[0] - br[0])**2 + (bl[1] -
br[1])**2)
                    heightLeft = np.sqrt((tl[0] - bl[0])**2 + (tl[1] - bl[1])**2)
                    heightRight = np.sqrt((tr[0] - br[0])**2 + (tr[1] -
br[1])**2)

                    maxWidth = int((widthTop+widthBottom)/2)
                    maxHeight = int((heightLeft+heightRight)/2)

                    dst = np.array([
                        [0,0],
                        [maxWidth-1, 0],
                        [maxWidth-1, maxHeight-1],
                        [0, maxHeight-1]], dtype='float32')

                    transformMatrix = cv2.getPerspectiveTransform(cornerslist,
dst)
                    scan = cv2.warpPerspective(img, transformMatrix, (maxWidth,
maxHeight))

                    kernel = np.array([[0,-1, 0],
                                       [-1, 5, -1],
                                       [0, -1, 0]])
                    scan = cv2.filter2D(src=scan, ddepth=-1, kernel=kernel)
                    #scan = cv2.cvtColor(scan, cv2.COLOR_BGR2GRAY)
                    #t = threshold_local(scan, 17, offset=15, method='gaussian')
                    #scan = (scan > t).astype('uint8') * 255

                    self.displayImage(img_track)
                    cv2.imshow("Captured Image", scan)
```

```
                    speech_function("Image captured")

                    cv2.imwrite('./Scanned/scanned.jpg', scan)
                    result = ocr_text('./Scanned/scanned.jpg')
                    print(result)
                    if len(result) == 0:
                        speech_function("Image empty or blurred. Please, press Q
to capture again")
                    else:
                        speech_function(result)
                    cv2.waitKey(0)
            else:

                # join stack images
                #imgStack = stack_images(0.5,
                                        #([img_track],
                                        #[np.zeros_like(img)]))

                #cv2.imshow('Processing Image', imgStack)
                self.displayImage(img_track)


            if cv2.waitKey(1) & 0xFF == ord('a'):
                break

    def displayImage(self, image):
        rgbImage = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)
        h, w, ch = rgbImage.shape
        bytesPerLine = ch * w
        convertToQtFormat = QImage(rgbImage.data, w, h, bytesPerLine,
QImage.Format_RGB888)
        p = convertToQtFormat.scaled(640, 480, Qt.KeepAspectRatio)
        return self.changePixmap.emit(p)

    def stop(self):
        self.ActiveThread = False
        self.quit()


class App(QMainWindow):
    def __init__(self):
        super().__init__()

        self.left=100
        self.top=100
```

```
        self.width= 660
        self.height= 500
        self.icon = './images/icon.jpg'
        self.initUI()

    @pyqtSlot(QImage)
    def setImage(self, image):
        self.label.setPixmap(QPixmap.fromImage(image))

    def initUI(self):
        self.setGeometry(self.left, self.top, self.width, self.height)
        self.setGeometry(self.left, self.top, self.width, self.height)
        self.setWindowTitle('Oculus OCR')
        self.setWindowIcon(QtGui.QIcon(self.icon))
        self.setMinimumSize(QtCore.QSize(self.width, self.height))
        self.setMaximumSize(QtCore.QSize(self.width, self.height))

        # create a label
        self.label = QLabel(self)
        self.label.move(10, 10)
        self.label.resize(imgWidth, imgHeight)

        # Loading the GIF
        self.movie = QMovie('./images/tech-eye.gif')
        self.movie.setScaledSize(QSize().scaled(imgWidth, imgHeight,
Qt.KeepAspectRatioByExpanding))
        self.label.setMovie(self.movie)
        self.startAnimation()

        # put the ThreadVision class into a ThreadVoice class to
        # use voice command on the App
        th = ThreadVision(self)
        th.changePixmap.connect(self.setImage)
        th.start()
        self.show()



    # Start Animation
    def startAnimation(self):
        self.movie.start()

    # Stop Animation(According to need)
    def stopAnimation(self):
        self.movie.stop()
```

```
stylesheet = """

    App {

        background-image: url("./images/background_image");
        background-repeat: no-repeat;
        background-position: center;
    }


"""
if __name__=='__main__':

    app = QApplication(sys.argv)
    dark_stylesheet = qdarkstyle.load_stylesheet_pyqt5()
    app.setStyleSheet(dark_stylesheet)
    window = App()
    window.show()
    sys.exit(app.exec_())
import cv2
import sys
from PyQt5.QtWidgets import  QWidget, QLabel, QApplication
from PyQt5.QtCore import QThread, Qt, pyqtSignal, pyqtSlot
from PyQt5.QtGui import QImage, QPixmap


webcam = True


class Thread(QThread):
    changePixmap = pyqtSignal(QImage)

    def run(self):

        if webcam:
            cap = cv2.VideoCapture(0)
        else:
            print("[INFO] camera not connected")

        self.ActiveThread = True
        while True:
            ret, frame = cap.read()
            if ret:
                # https://stackoverflow.com/a/55468544/6622587
                rgbImage = cv2.cvtColor(frame, cv2.COLOR_BGR2RGB)
```

```
                h, w, ch = rgbImage.shape
                bytesPerLine = ch * w
                convertToQtFormat = QImage(rgbImage.data, w, h, bytesPerLine,
QImage.Format_RGB888)
                p = convertToQtFormat.scaled(640, 480, Qt.KeepAspectRatio)
                self.changePixmap.emit(p)

    def stop(self):
        self.ActiveThread = False
        self.quit()


class App(QWidget):
    def __init__(self):
        super().__init__()
        [...]
        self.initUI()

    @pyqtSlot(QImage)
    def setImage(self, image):
        self.label.setPixmap(QPixmap.fromImage(image))

    def initUI(self):
        self.setWindowTitle(self.title)
        self.setGeometry(self.left, self.top, self.width, self.height)
        self.resize(1800, 1200)
        # create a label
        self.label = QLabel(self)
        self.label.move(280, 120)
        self.label.resize(640, 480)
        th = Thread(self)
        th.changePixmap.connect(self.setImage)
        th.start()
        self.show()

def stack_images(scale, img_array):

    rows = len(img_array)
    cols = len(img_array[0])
    rows_available = isinstance(img_array[0], list)
    width = img_array[0][0].shape[1]
    height = img_array[0][0].shape[0]
    if rows_available:
        for x in range(0, rows):
            for y in range(0, cols):
```

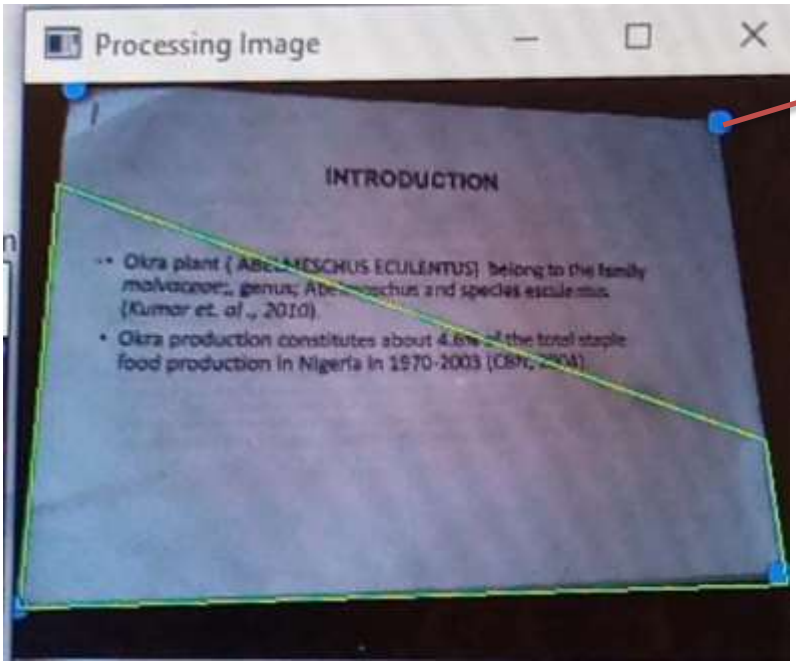
```
                if img_array[x][y].shape[:2] == img_array[0][0].shape[:2]:
                    img_array[x][y] = cv2.resize(img_array[x][y], (0, 0), None,
scale, scale)
                else:
                    img_array[x][y] = cv2.resize(img_array[x][y],
(img_array[0][0].shape[1], img_array[0][0].shape[0]),
                                                 None, scale, scale)
                if len(img_array[x][y].shape) == 2: img_array[x][y] =
cv2.cvtColor(img_array[x][y], cv2.COLOR_GRAY2BGR)
        image_blank = np.zeros((height, width, 3), np.uint8)
        hor = [image_blank] * rows
        hor_con = [image_blank] * rows
        for x in range(0, rows):
            hor[x] = np.hstack(img_array[x])
        ver = np.vstack(hor)
    else:
        for x in range(0, rows):
            if img_array[x].shape[:2] == img_array[0].shape[:2]:
                img_array[x] = cv2.resize(img_array[x], (0, 0), None, scale,
scale)
            else:
                img_array[x] = cv2.resize(img_array[x], (img_array[0].shape[1],
img_array[0].shape[0]), None, scale,
                                          scale)
            if len(img_array[x].shape) == 2: img_array[x] =
cv2.cvtColor(img_array[x], cv2.COLOR_GRAY2BGR)
        hor = np.hstack(img_array)
        ver = hor

    return ver
```

APPENDIX III

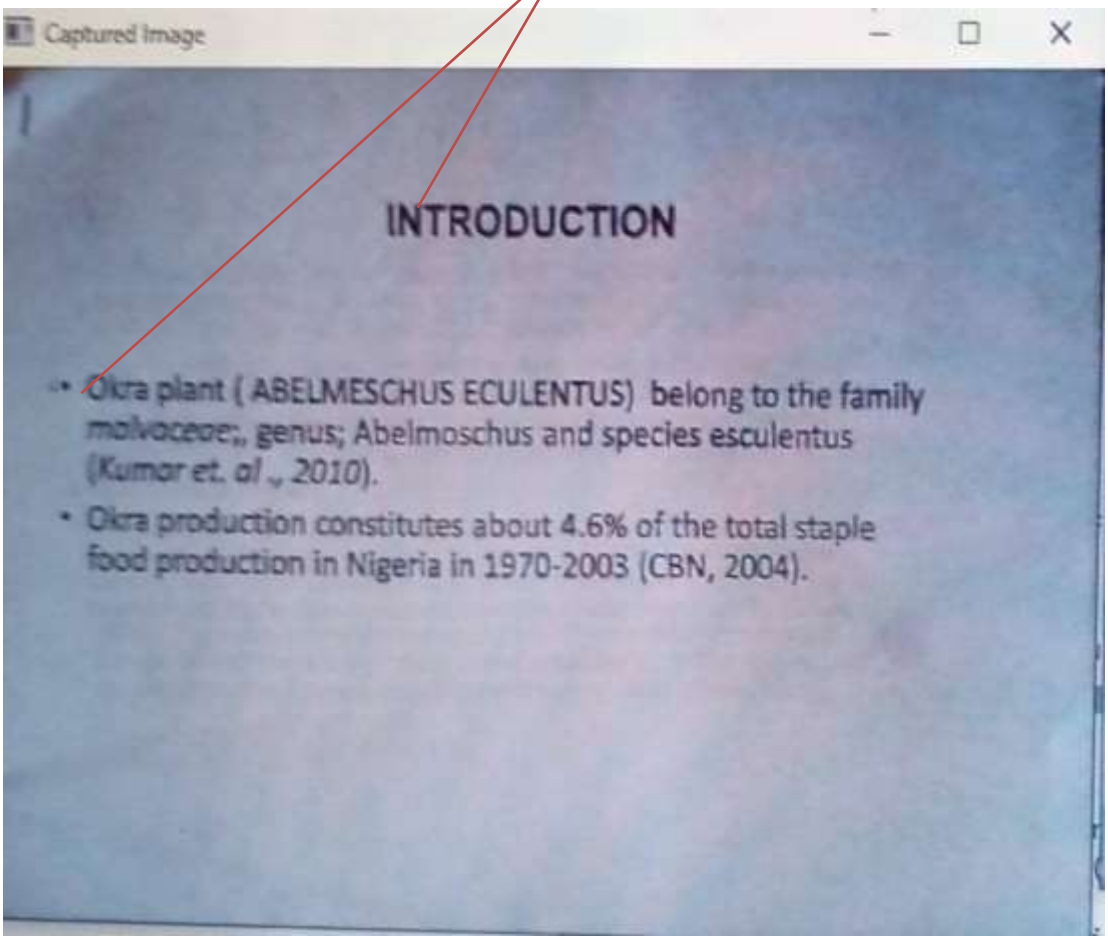
SAMPLE OUTPUTS

System's Interface



System Reading the edges of document before capturing

Detected and recognized text ready for audio output



APPENDIX C

Sample Of Synth90k Data Set

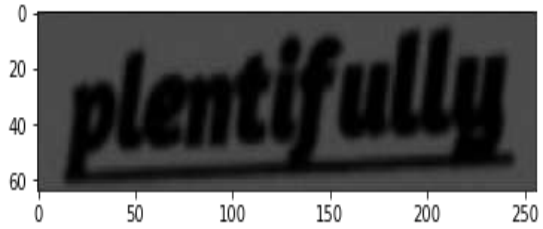


Figure 4.9a: Test data1 Original: PLENTIFULLY
Predicted: PLENTIFULLY

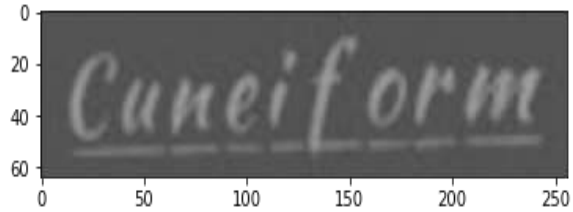


Figure 4.9b: Test data2 Original: CUNEIFORM
Predicted: CUNEIFORM

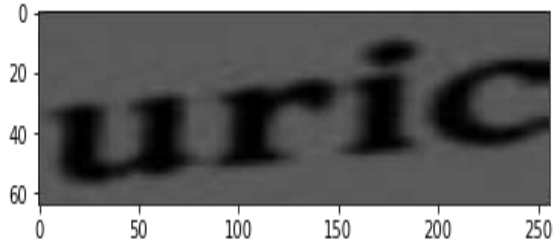


Figure 4.9c: Test data3 Original: URIC
Predicted: URIC

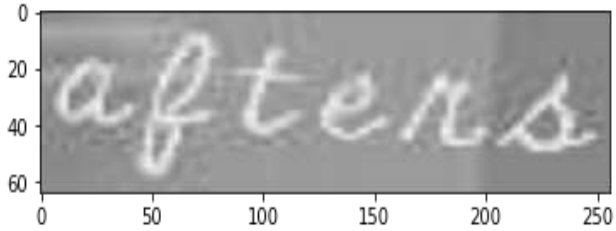


Figure 4.9d: Test data4 Original: AFTERS
predicted: AFTERS

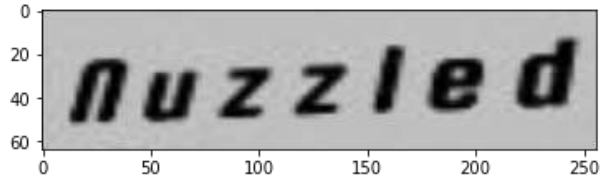


Figure 4.9e: Test data5 Original: NUZZLED
Predicted: NUZZLED

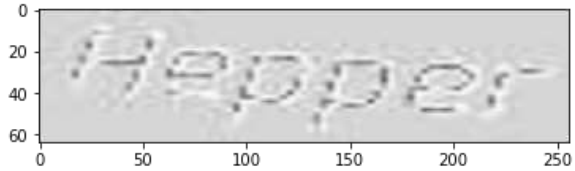


Figure 4.9f: Test data6 Original: HEPPER
Predicted: HEPPE

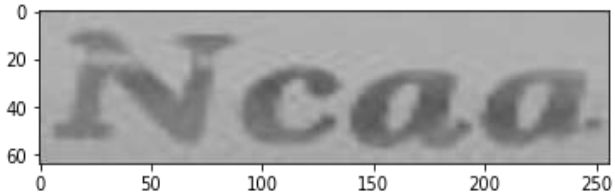


Figure 4.9g: Test data7 Original: NCAA
Predicted: NCAA

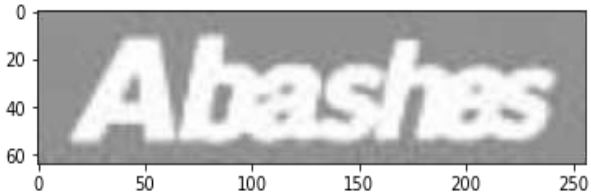


Figure 4.9h: Test data8 Original: ABASHES
Predicted: ABASHES

Figure 4.9i: Test data 9 **Original: WINGDINGS**
Predicted: WINGDINGS

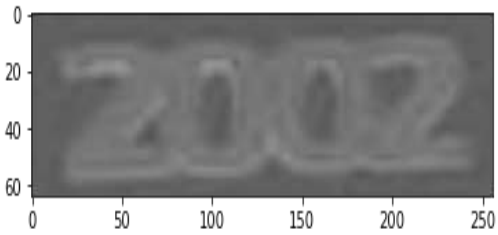


figure 4.9j: Test data10 **Original: 2002**
Predicted: Z002

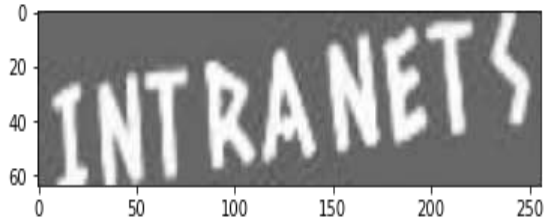


Figure 4.9k: Test data10 **Original: INTRANETS**
Predicted: INTRANET2

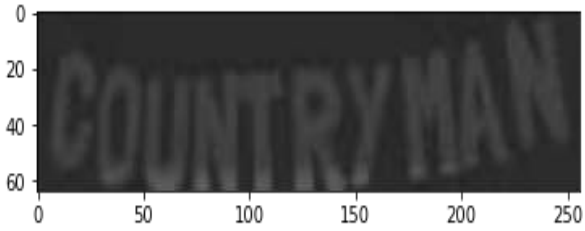


figure 4.9l: Test data 11 Original: COUNTRYMAN
Predicted: COUNTRYMAN

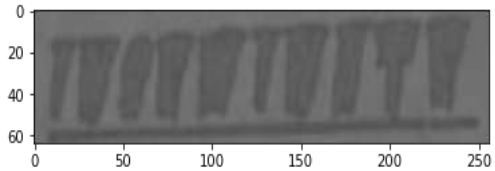


Figure 4.9m: Test data12 **Original: INSEMINATE**
Predicted: TARRIAIIA

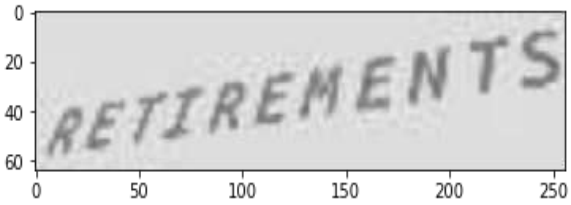


figure 4.9n: Test Data13 **Original: RETIREMENTS**
Predicted: RETTREMENTS

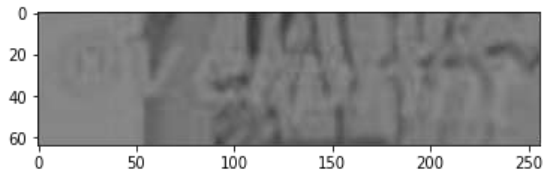


Figure 4.9o: Test Data14 **Original: OVERPRINT**
Predicted: OVERRIIT

figure 4.9p: Test data15 **Original: ANTHROPOMORPHIC**
Predicted: INTEHOOPOPPHI