“Image To Speech Converter”

Submitted in partial fulfilment of the requirements
of the degree of

Bachelor of Engineering

by

Uddesh Karda

Vaibhav Gaikwad

Aniket Nighot

under the guidance of

Supervisor:

Mrs.Asma Parveen I. Siddavtam

Asst. Prof., Dept of Information Technology

V.E.S.

Since 1962

Department of Information Technology

Vivekanand Education Society’s Institute of Technology

2017-18

PROJECT REPORT APPROVAL FOR B. E.

This project report entitled Image to Speech converter by **Uddesh Karda, Vaibhav Gaikwad** and **Aniket Nighot** is approved for the degree of **B.E. (Discipline of Information Technology)**

Examiners:

1) ______________________________

2) ______________________________

Supervisors:

1) ______________________________

Date:

Place:

DECLARATION

We declare that this written submission represents our ideas in our own words and where other ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. We understand that any violation of the above will cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

____

(Signature)

Uddesh Karda – D20/32

____

(Signature)

Aniket Nighot – D20/49

____

(Signature)

Vaibhav Gaikwad – D20/16

ABSTRACT

There are about 285 million people who are visually impaired worldwide. Most humans tend to read books, newspapers, magazines, blogs, etc to gain knowledge. However sometimes it becomes impossible to read (example- travelling). Several efforts have been made in order to give the visually impaired people access to information such as newspapers, magazines etc. One of the solution is transforming textual information into speech information.

The aim of this project is to convert text in an image, taken by the user's smartphone camera into speech with increased computation speed and also keeping a high accuracy rate. The technology that allows us to convert text in images captured by an input device into an editable, searchable Further this text generated can be converted into speech using the inbuilt android libraries of text to speech conversion.

Keywords: OCR,Image to Text,Image to Speech,Image Processing.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Real world contains too many significant message and useful information cannot be ignored or left unread. Sometimes a signboard or any other notice could carry an important message or even danger notice that could be missed by visually impaired people. This application is mainly beneficial for visually impaired people to access printed text which may carry significant messages.

If the message is unreachable to mankind either due to biological barriers or linguistic barriers it might cause important information to be missed out which could lead to harm. Therefore this application will also be useful to the travellers and tourists, students, illiterate people to overcome the language barrier.

1.2. Problem Statement

To develop an android app for visually impaired people or people with biological or linguistic barriers which will help them to access the information around them in printed format by converting text in an image into speech.

1.3. Objectives

To convert text in an image, taken by the user's smartphone camera into speech with increased computation speed and also keeping a high accuracy rate. Also user can upload a pdf file which would be converted into speech.

Chapter 2

Literature Survey

2.1 Literature/Papers Studied

2.1.1 Detecting Text Based Image With Optical Character Recognition for English Translation and Speech using Android

Abstract:-

Smartphones have been known as most commonly used electronic devices in daily life today. As hardware embedded in smartphones can perform much more task than traditional phones, the smartphones are no longer just a communication device but also considered as a powerful computing device which able to capture images, record videos, surf the internet and etc. With advancement of technology, it is possible to apply some techniques to perform text detection and translation. Therefore, an application that allows smartphones to capture an image and extract the text from it to translate into English and speech it out is no longer a dream. In this study, an Android application is developed by integrating Tesseract OCR engine, Bing translator and phones' built-in speech out technology. Final deliverable is tested by various type of target end user from a different language background and concluded that the application benefits many users. By using this app, travelers who visit a foreign country able to understand messages portrayed in different language. Visually impaired users are also able to access important message from a printed text through speech out feature.

2.1.2 Optical Character Recognition (OCR) Performance in Server-based Mobile Environment

Abstract:-

There are several Optical Character Recognition (OCR) mobile applications on the market running on mobile devices, both android and iOS (iPhone, iPad, iPod) platforms. The limitations of mobile device processor hinder the possible execution of computationally intensive applications that need less time of process. This paper proposes a framework of Optical Character Recognition (OCR) on mobile device using server-based processing. Comparison methods proposed by this paper by conducting a series of tests using standalone and server-based OCR on mobile devices, and compare the results of the accuracy and time required for the entire OCR processing. Server-based mobile OCR obtains 5% higher character recognition accuracy than the standalone OCR and its format recognition accuracy is 99.8%. The framework tries to overcome the limitation of mobile device capability process, so the devices can do the computationally intensive application more quickly.

2.1.3 Medical Document Reader on Android Smartphone

Abstract:-

This paper presents a method for reading medical documents by using an Android smartphone. We have used techniques based on the Tesseract OCR Engine to extract the text content from medical document images such as a physical examination report. The following factors related to the document are considered: character font, text block size, and distance between the document and the camera on the phone. Based on experimental results, we found that among three character fonts (Angsana New, Calibri, and Tahoma), Calibri and Tahoma gave very high average accuracies (greater than 90%) for both character recognition and word recognition, but Angsana New gave quite a lower accuracy, about 75%. For the optimal distance between the document and the smartphone, the recommended distance is from 12 cm. to 15 cm. for a document block size of 21 x3, 13 x 10, 12 x 8, or 10 x 13 cmz.

2.1.4 Proposal for Automatic License and Number Plate Recognition System for Vehicle Identification

Abstract:-

In this paper, we propose an automatic and mechanized license and number plate recognition (LNPR) system which can extract the license plate number of the vehicles passing through a given location using image processing algorithms. No additional devices such as GPS or radio frequency identification (RFID) need to be installed for implementing the proposed system. Using special cameras, the system takes pictures from each passing vehicle and forwards the image to the computer for being processed by the LPR software. Plate recognition software uses different algorithms such as localization, orientation, normalization, segmentation and finally optical character recognition (OCR). The resulting data is applied to compare with the records on a database. Experimental results reveal that the presented system successfully detects and recognizes the vehicle number plate on real images. This system can also be used for security and traffic control.

2.1.5 Optical Character Recognition Technique Algorithms

Abstract:-

In this paper, we present a new neural network (NN) based method for optical character recognition (OCR) as well as handwritten character recognition (HCR). Experimental results show that our proposed method achieves increased accuracy in optical character recognition as well as handwritten character recognition. We present through an overview of existing handwritten character recognition techniques. All the algorithms describes more or less on their own. Handwritten character recognition is a very popular and computationally expensive task; we describe advanced approaches for handwritten character recognition. In the present work, we would like to compare the most important once out of the variety of advanced existing techniques, and we will systematize the techniques by their characteristic considerations. It leads to the behaviour of the algorithms reaches to the expected similarities.

Chapter 3

Requirements and Analysis

3.1. Functional and Non Functional Requirements

Functional Requirements

- Software should process the image ,extract the characters and give output in the form of speech.
- Software should provide a way to load scanned document for conversion purpose.
- Software should be able to translate text to user specified language.

Non-Functional Requirements

- **Accuracy :** Extent to which software satisfies its specifications and fulfills the objective.
- **Modifiability:** Requirements about the effort required to make changes in the software. Often, the measurement is personnel effort (person- months).
- **Speed :** Time required for the software to complete its task
- **Usability:** This requirement specifies the level of sufficiency and operability by the end users of the system .It features the level of difficulty to learn and operate the system. The

requirements are often expressed in effective knowledge gain per person or similar metrics

3.2. Constraints

- **Platform constraints:** These constraints ensure the discussion of the platform dependent and independent parameters of the system. This parameter is specified as per the user requests. Various modifications in the features of platform are possible

3.3. Hardware and Software Requirements

3.3.1 Hardware Requirements

- **Android Mobile Device with auto-focus enabled camera with 3G data connectivity :** It lets users take photo. Auto focus feature enable to capture photos with good quality.
- **Server :-** Image processing is carried out at server side to provide frequent results to the users.

3.3.2 Software Requirements

- **Android Studio for Android application development:** Android Studio provides the fastest tools for building apps on every type of Android device. World-class code editing, debugging, performance tooling, a flexible build system, and an instant build/deploy system all allow you to focus on building unique and high quality apps.

- **Yandex translator API:** The API provides access to the Yandex online machine translation service. It supports more than 90 languages and can translate separate words or complete texts. The API makes it possible to embed Yandex.Translate in a mobile app or web service for end users. Or translate large quantities of text, such as technical documentation.

- **Python OpenCV:** It is a library of many inbuilt functions mainly aimed at real time image processing. Now it has several hundreds of image processing and computer vision algorithms which make developing advanced computer vision applications easy and efficient. Optimized for real time image processing & computer vision applications.

- **Tesseract OCR libraries by Google:** Tesseract is an optical character recognition engine for various operating systems. It is free software, released under the Apache License, Version 2.0, and development has been sponsored by Google since 2006. In 2006 Tesseract was considered one of the most accurate open-source OCR engines then available.

Chapter 4

Proposed Design

4.1 System Architecture

The system architecture consists of two major modules:

A. Mobile Device:

Under the mobile device we have the following main components:

1. **Camera** : The application layer camera software which will provide the video frames.
2. **HTTP web request**: It represents the HTTP request object in Java which will encapsulate the data to be sent to the server.
3. **Text To Speech (TTS) Engine:**TTS engine is needed to give speech output to user

B. Server for online computation:

1. **Remote OCR engine**: This engine provides online OCR facility with faster computation.
2. **Remote Translate Engine :**This engine provides online translation facility.

Following diagram shows the architecture of the Image to speech converter:

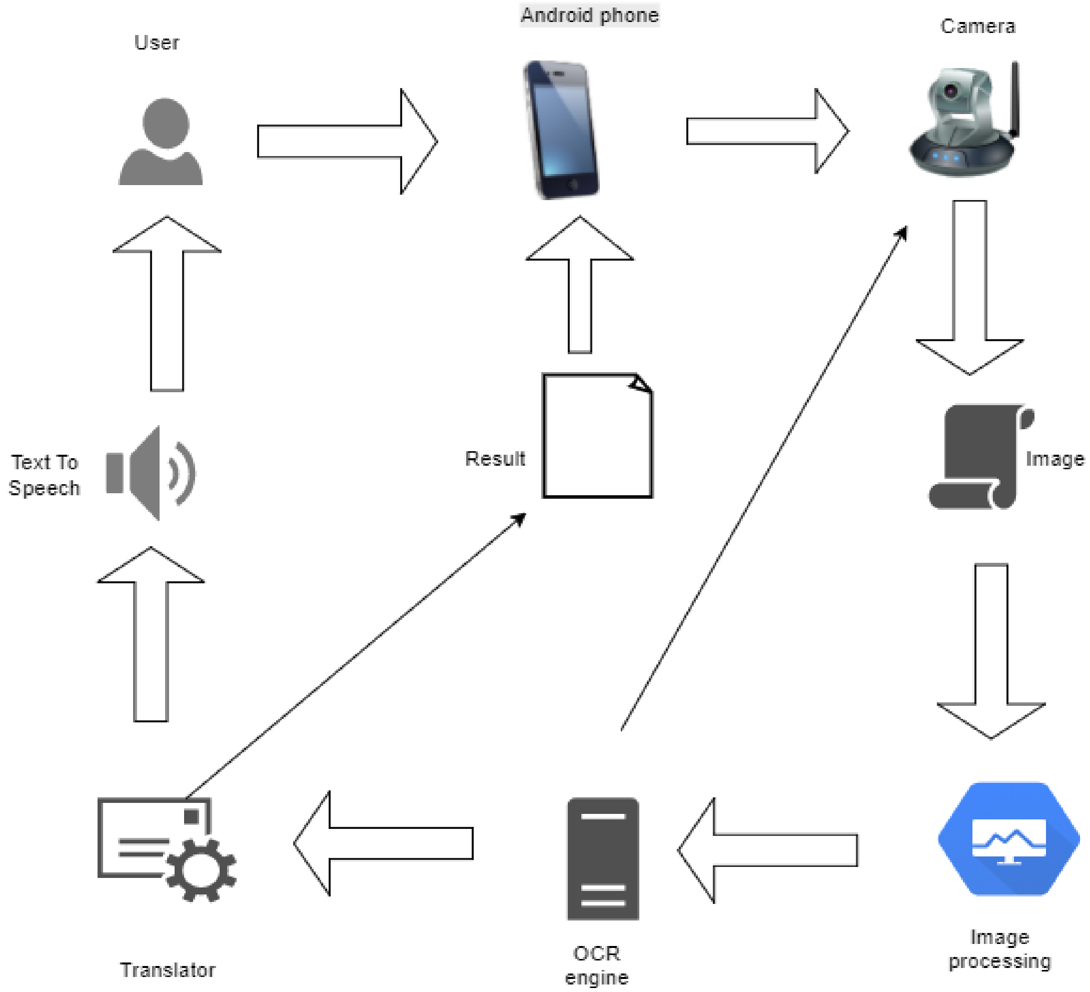


Figure 4.1 System Architecture

4.2 Detailed Design

A. Flowchart

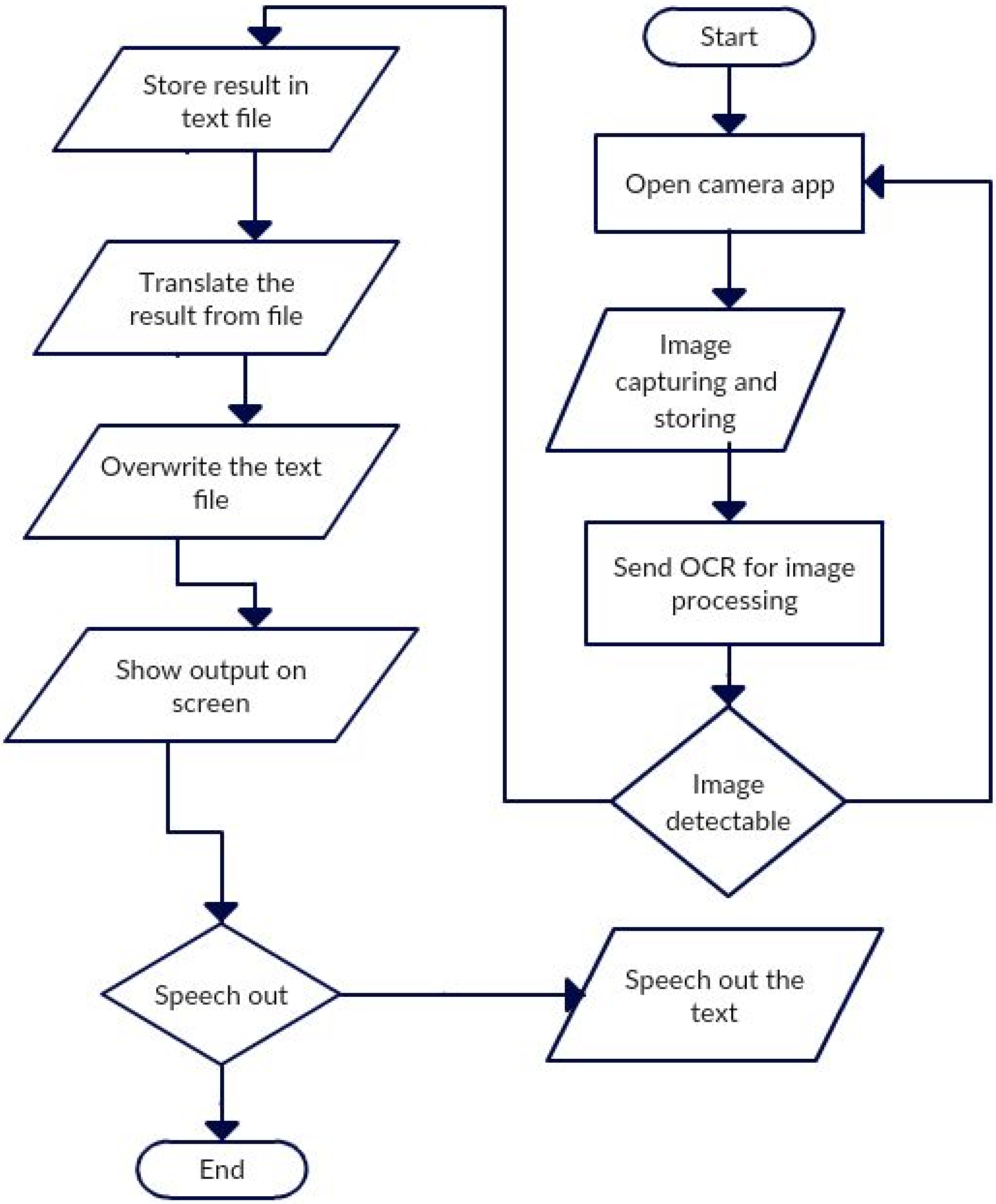


Figure.4.2 Flowchart of Image to speech converter system

4.2.1 Image capturing and pre-processing

Firstly,images are captured or loaded from mobile device.This images may contain noise which affects output of the OCR. Therefore there are some techniques such as Image filtering for noise reduction,binarization,text segmentation to be done in the preprocessing phase to improve performance and accuracy of the character recognition system.

4.2.2.1 Scaling

Image is scaled to the right size which usually is of at least 300 DPI (Dots Per Inch). Keeping DPI lower than 200 gives unclear and incomprehensible results while keeping the DPI above 600 unnecessarily increase the size of the output file without improving the quality of the file. Thus, a DPI of 300 works best for this purpose.

4.2.2.2 Image Segmentation

In computer vision, segmentation is the process of partitioning a digital image into multiple segments (sets of pixels).In OCR preprocessing image segmentation is used to detect and separate text content from the entire image.

4.2.2.3 Image Filtering and Noise Reduction

In image processing, filters are mainly used to suppress either the high frequencies in the image, i.e. smoothing the image, or the low frequencies, Image restoration and enhancement techniques are described in both the spatial domain and frequency domain, i.e. Fourier transforms. However, Fourier transforms require substantial computations, and in some cases are not worth the effort.Using a small convolution mask, such as 3x3, and convolving this mask over an image is much easier and faster than performing.Fourier transforms and multiplication; therefore, only spatial filtering techniques are used.Widely used image filtering techniques are Gaussian filtering and Median filtering.

OpenCV (Open Source Computer Vision Library) is an open source computer vision and machine learning software library. OpenCV was built to provide a common infrastructure for computer vision applications hence it is used to do image preprocessing for OCR.OpenCV is released under a BSD license and hence it’s free for both academic and commercial use. It has C++, Python and Java interfaces and supports Windows, Linux, Mac OS, iOS and Android.

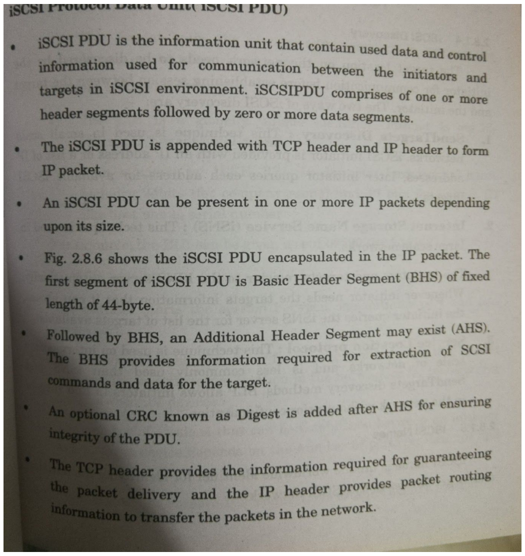
4.2.2.3 Binarization or Thresholding

Binarization is the process of converting a pixel image to a binary image. In character recognition systems most of the applications binary images since processing colour images is computationally high..Hence image thresholding is performed to convert coloured image into binary image.

The simplest thresholding methods replace each pixel in an image with a black pixel if the image intensity I I(i,j) is less than some fixed constant T that is,I(i,j) < T or a white pixel if the image intensity is greater than that constant.

But images generally has different lighting conditions in different areas. Hence instead of global thresholding(Fix threshold value for all pixels in an image) we go for adaptive thresholding.

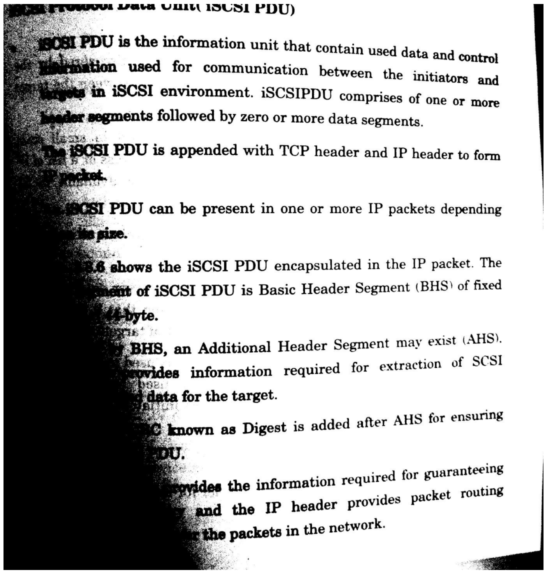
In adaptive threshold unlike fixed or global threshold, the threshold value at each pixel location depends on the neighboring pixel intensities



iSCSI Protocol Data Unit(iSCSI PDU)

- iSCSI PDU is the information unit that contain used data and control information used for communication between the initiators and targets in iSCSI environment. iSCSIPDU comprises of one or more header segments followed by zero or more data segments.
- The iSCSI PDU is appended with TCP header and IP header to form IP packet.
- An iSCSI PDU can be present in one or more IP packets depending upon its size.
- Fig. 2.8.6 shows the iSCSI PDU encapsulated in the IP packet. The first segment of iSCSI PDU is Basic Header Segment (BHS) of fixed length of 44-byte.
- Followed by BHS, an Additional Header Segment may exist (AHS). The BHS provides information required for extraction of SCSI commands and data for the target.
- An optional CRC known as Digest is added after AHS for ensuring integrity of the PDU.
- The TCP header provides the information required for guaranteeing the packet delivery and the IP header provides packet routing information to transfer the packets in the network.

Figure 4.3 Input Image



Data Unit(iSCSI PDU)

iSCSI PDU is the information unit that contain used data and control
information used for communication between the initiators and
targets in iSCSI environment. iSCSIPDU comprises of one or more
header segments followed by zero or more data segments.

The iSCSI PDU is appended with TCP header and IP header to form
IP packet.

SCSI PDU can be present in one or more IP packets depending
its size.

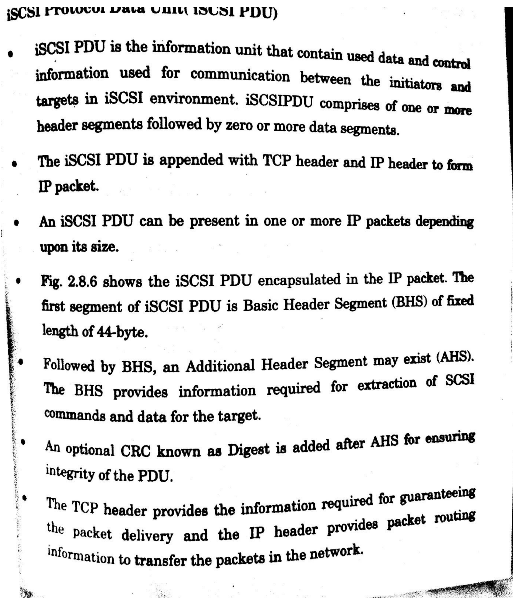
.6 shows the iSCSI PDU encapsulated in the IP packet. The
of iSCSI PDU is Basic Header Segment (BHS) of fixed
-byte.

BHS, an Additional Header Segment may exist (AHS).
provides information required for extraction of SCSI
data for the target.

C known as Digest is added after AHS for ensuring
DU.

provides the information required for guaranteeing
and the IP header provides packet routing
the packets in the network.

Figure 4.4 After Fixed Thresholding



iSCSI Protocol Data Unit(iSCSI PDU)

- iSCSI PDU is the information unit that contain used data and control information used for communication between the initiators and targets in iSCSI environment. iSCSIPDU comprises of one or more header segments followed by zero or more data segments.
- The iSCSI PDU is appended with TCP header and IP header to form IP packet.
- An iSCSI PDU can be present in one or more IP packets depending upon its size.
- Fig. 2.8.6 shows the iSCSI PDU encapsulated in the IP packet. The first segment of iSCSI PDU is Basic Header Segment (BHS) of fixed length of 44-byte.
- Followed by BHS, an Additional Header Segment may exist (AHS). The BHS provides information required for extraction of SCSI commands and data for the target.
- An optional CRC known as Digest is added after AHS for ensuring integrity of the PDU.
- The TCP header provides the information required for guaranteeing the packet delivery and the IP header provides packet routing information to transfer the packets in the network.

Figure 4.5 After Local Adaptive Thresholding

Figure 4.5 shows the final preprocessed image which is passed to OCR engine for text extraction

4.2.2 OCR

There are numerous open sources as well as commercial OCR engines available in the market with their own strengths and weaknesses. Many open source communities offer engines such as GOCR, Cuneiform, OCRAD, Tesseract and OCROPUS. There are commercially available OCR engines such as ABBYY Finereader, OmniPage, and Microsoft Office Document Imaging.

Tesseract works well on all computer operating system as well as Android and iPhone mobile platform. Due to popularity of Tesseract being open source engine, there are a lot of academic experiments and OCR software developments conducted successfully. Based on study conducted between OCRAD, GOCR and Tesseract, found out that the Tesseract outperform other open source engines. Despite of unclean data, Tesseract proved the best free and open source OCR engine in term of accuracy and processing time as shown in below table :

Parameter	Cuneiform	GOCR	OCRD	Tesseract
license	BSD	GPL2	GPL3	Apache 2.0
courier/black	61%	67%	21%	81%
courier/gray	×	67%	21%	81%
times/black	94%	76%	82%	92%
times/gray	×	76%	82%	92%
verdana/black	95%	97%	97%	95%
justy/black	3%	31%	1%	15%
justy/gray	×	31%	15%	15%

Table 1. Comparison between various OCR's

4.2.2 Translation

There are various translators available in the market such as Google Translator, Bing Translator, etc. Google Translator and Bing Translators are paid API's and provide access to only one user per access key.

In this proposed system Yandex translator will be used.Yandex is a translator that provides free API and multiple client access using one key. It also provides synchronized translation for 95 languages, predictive typing, dictionary with transcription, pronunciation and usage examples, and many other features.

4.2.3 Text To Speech

For text to speech, phone built-in feature would perform the speech out service. Android libraries such as android.text and android.speech will be used mainly for this purpose.Other available options were espeak, live-text-view and AndroidMary-TTS but the best option was the inbuilt text to speech libraries provided by google itself.

Chapter 5

Implementation

5.1 Technologies Used

1. Java
2. Android
3. PHP
4. Python
5. XML
6. JSON
7. Tesseract OCR Engine
8. OpenCV
9. Yandex Translator API

5.2 Code Snippets

5.2.1 Image Scaling

```
def set_image_dpi(file_path):
    im = Image.open(file_path)
    length_x, width_y = im.size
    factor = max(1, int(IMAGE_SIZE / length_x))
    size = factor * length_x, factor * width_y
    # size = (1800, 1800)
    im_resized = im.resize(size, Image.ANTIALIAS)
    temp_file = tempfile.NamedTemporaryFile(delete=False, suffix='.jpg')
    temp_filename = temp_file.name
    im_resized.save(temp_filename, dpi=(300, 300))
    return temp_filename
```

5.2.2 Image Segmentation

```
def segmentation(filepath):
    image = cv2.imread(filepath)
    gray = cv2.cvtColor(image,cv2.COLOR_BGR2GRAY) # grayscale
    _,thresh = cv2.threshold(gray,150,255,cv2.THRESH_BINARY_INV) # threshold
    kernel = cv2.getStructuringElement(cv2.MORPH_CROSS,(3,3))
    dilated = cv2.dilate(thresh,kernel,iterations = 13) # dilate
    _, contours, hierarchy =         cv2.findContours(dilated,cv2.RETR_EXTERNAL,
    cv2.CHAIN_APPROX_NONE)
    # get contours
    text=""
    # for each contour found, draw a rectangle around it on original image
    for contour in contours:
        # get rectangle bounding contour
        [x,y,w,h] = cv2.boundingRect(contour)
        # discard areas that are too large
        if h>300 and w>300:
            continue
        # discard areas that are too small
        if h<40 or w<40:
            continue
        # draw rectangle around contour on original image
        cv2.rectangle(image,(x,y),(x+w,y+h),(255,0,255),2)
        cropped = image[y :y +  h , x : x + w]
        s =  'mysite/images/s.jpg'
        cv2.imwrite(s , cropped)
        return(s)
```

5.2.3 Image Smoothening and Thresholding

```
def thresholding(filepath):
        image=cv2.imread(filepath)
        blur2 = cv2.GaussianBlur(image, (1,1), 0)
        th =cv2.adaptiveThreshold(blur2,255,cv2.ADAPTIVE_THRESH_GAUSSIAN_C
        ,cv2.THRESH_BINARY,513,15)
        median = cv2.medianBlur(th,1)
        ret2, th2 = cv2.threshold(median, 0, 255 ,   cv2. THRESH_BINARY+ cv2. THRESH_
        OTSU)
        cv2.imwrite("mysite/images/final.png",th2)
        return(median)
```

5.2.3 Using Pytesseract to Extract text

```
text = pytesseract.image_to_string(Image.open(filepath))
```

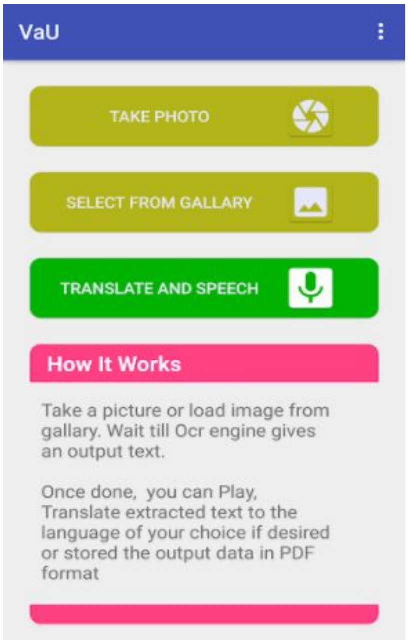
Chapter 6

Results

6.1 SCREENSHOTS

These are the screenshots of the result

A. Home Screen :

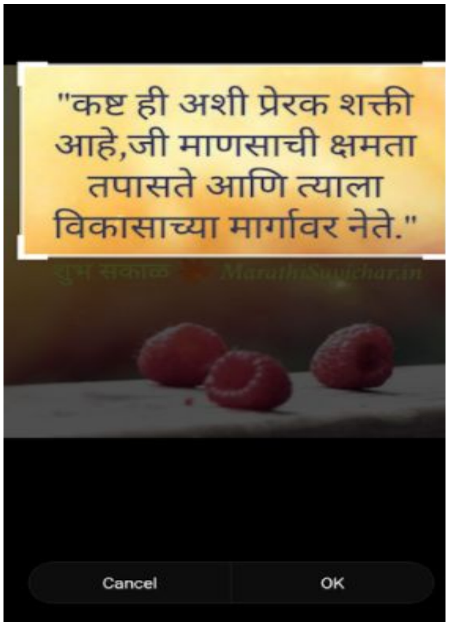


Take a picture or load image from gallary. Wait till Ocr engine gives an output text.

Once done, you can Play, Translate extracted text to the language of your choice if desired or stored the output data in PDF format

Figure 6.1 Screenshot of Homescreen

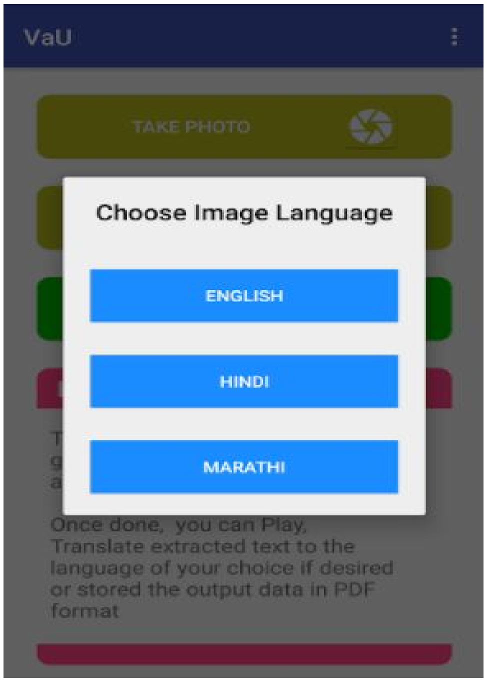
B. Image Cropping in app :



"कष्ट ही अशी प्रेरक शक्ती
आहे,जी माणसाची क्षमता
तपासते आणि त्याला
विकासाच्या मार्गावर नेते."

Figure 6.2 Screenshot of image cropping

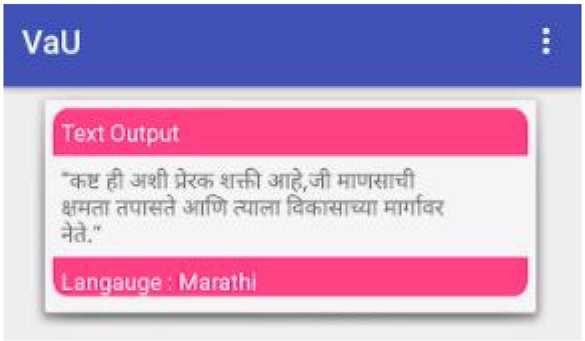
C.Select Language :



Once done, you can Play, Translate extracted text to the language of your choice if desired or stored the output data in PDF format

Figure 6.3 Selecting Language

D.Output text from server:

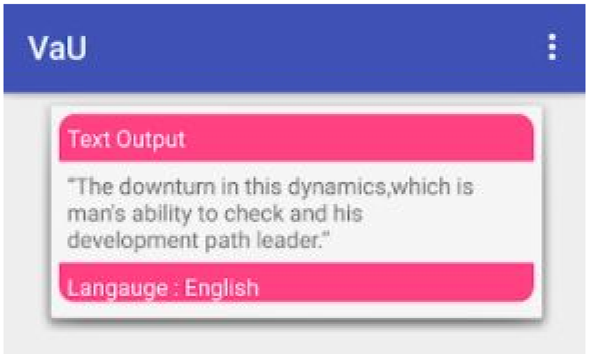


"कष्ट ही अशी प्रेरक शक्ती आहे,जी माणसाची क्षमता तपासते आणि त्याला विकासाच्या मार्गावर नेते."

Langauge : Marathi

Figure 6.4 Screenshot of Text output

E.Translator output:

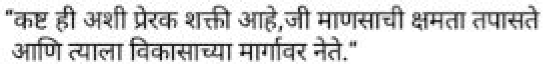


"The downturn in this dynamics,which is man's ability to check and his development path leader."

Langauge : English

Figure 6.5 Screenshot of Translated text

F.Generated Pdf Output:



"कष्ट ही अशी प्रेरक शक्ती आहे,जी माणसाची क्षमता तपासते आणि त्याला विकासाच्या मार्गावर नेते."

Figure 6.6 Screenshot of generated PDF

Chapter 7

Conclusion

Chapter 8

Future Scope

- This application can be extended to detect scientific and mathematical printed text and convert it into digital format E.g. Pdf, text file
- Many other languages can be added in the application.

Chapter 9

References

[1] Canedo-Rodriguez, S. Kim, J. Kim and Y. Blanco-Fernandez, 'English to Spanish translation of signboard images from mobile phone camera', IEEE Southeastcon 2009, 2009.

[2] OCR for Mobile Phones Kathryn Hymes and John Lewin

[3] Teddy Mantoro, Abdul Muis Sobri, Wendi Usino, "Optical Character Recognition (OCR) Performance in Server-based Mobile Environment", 2013 International Conference on Advanced Computer Science Applications and Technologies

[4] 2016 1st International Conference on New Research Achievements in Electrical and Computer Engineering Proposal for "Automatic License and Number Plate Recognition System for Vehicle Identification" by Hamed Saghaei

Documentation:

[5] OpenCV Documentation- https://docs.opencv.org/2.4/doc/tutorials/tutorials.html